LECTURE NOTES IN COMPUTATIONAL SCIENCE AND ENGINEERING 112

Bülent Karasözen · Murat Manguoğlu
Münevver Tezer-Sezgin · Serdar Göktepe
Ömür Uğur *Editors*

Numerical Mathematics and Advanced Applications ENUMATH 2015

Springer

Lecture Notes in Computational Science and Engineering

112

More information about this series at http://www.springer.com/series/3527

Bülent Karasözen • Murat Manguoğlu •
Münevver Tezer-Sezgin • Serdar Göktepe •
Ömür Uğur
Editors

Numerical Mathematics and Advanced Applications ENUMATH 2015

Springer

Editors

Bülent Karasözen
Mathematics & Applied Mathematics
Middle East Technical University
Ankara, Turkey

Murat Manguoğlu
Department of Computer Engineering
Institute of Applied Mathematics
Ankara, Turkey

Münevver Tezer-Sezgin
Department of Mathematics
Middle East Technical University
Ankara, Turkey

Serdar Göktepe
Civil Engineering & Applied Mathematics
Middle East Technical University
Ankara, Turkey

Ömür Uğur
Institute of Applied Mathematics
Middle East Technical University
Ankara, Turkey

ISSN 1439-7358 ISSN 2197-7100 (electronic)
Lecture Notes in Computational Science and Engineering
ISBN 978-3-319-39927-0 ISBN 978-3-319-39929-4 (eBook)
DOI 10.1007/978-3-319-39929-4

Library of Congress Control Number: 2016956100

Mathematics Subject Classification (2010): 65XX, 74XX, 76XX, 78XX

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

Preface

The European Conference on Numerical Mathematics and Advanced Applications (ENUMATH) is a series of conferences held every 2 years to provide a forum for discussion on recent aspects of numerical mathematics and scientific and industrial applications. The previous ENUMATH meetings took place in Paris (1995), Heidelberg (1997), Jyvaskyla (1999), Ischia (2001), Prague (2003), Santiago de Compostela (2005), Graz (2007), Uppsala (2009), Leicester (2011), and Lausanne (2013).

This book contains a selection of invited and contributed lectures of the ENUMATH 2015 organized by the Institute of Applied Mathematics, Middle East Technical University, Ankara, Turkey, September 14–18, 2015. It gives an overview of recent developments in numerical analysis, computational mathematics, and applications by leading experts in the field. The conference attracted around 300 participants from around the world including 11 invited talks by:

- Assyr Abdulle (EPF Lausanne, Switzerland), *Reduced Basis Multiscale Methods*
- Rémi Abgrall (Universität Zürich, Switzerland), *Recent Progress on Non-Oscillatory Finite Element Methods for Convection Dominated Problems*
- Burak Aksoylu (TOBB University of Economics and Technology, Ankara, Turkey), *Incorporating Local Boundary Conditions into Nonlocal Theories*
- Mark Ainsworth (Brown University, Providence, USA), *Multigrid at Scale?*
- Willi Freeden (TU Kaiserslautern, Germany), *Principles in Geomathematically Reflected Numerics and Their Application to Inverse Potential Methods in Geothermal Exploration*
- Des Higham (University of Strathclyde, Glasgow, UK), *Models and Algorithms for Dynamic Networks*
- Yvon Maday (UniversitéĄ Pierre et Marie Curie, Paris, France), *Towards a Fully Scalable Balanced Parareal Method: Application to Neutronics*
- Kaisa Miettinen (University of Jyväskylä, Finland), *Examples of Latest Interactive Method Developments in Multiobjective Optimization*
- Mario Ohlberger (Universität Münster, Germany), *Localized Model Reduction for Multiscale Problems*

- Anders Szepessy (KTH, Stockholm, Sweden), *On Global and Local Error with Application to Adaptivity, Inverse Problems and Modeling Error*
- Eugene E. Tyrtyshnikov (Russian Academy of Sciences, Moscow, Russia), *Tensor Decompositions and Low-Rank Matrices in Mathematics and Applications*

There were 119 minisymposia presentations in 20 sessions, and 89 contributed talks covering a broad spectrum of numerical mathematics. This ENUMATH 2015 proceeding will be useful for a wide range of readers giving them a state-of-the-art overview of advanced techniques, algorithms, and results in numerical mathematics and scientific computing. This book contains a selection of 61 papers by the invited speakers and from the minisymposia as well as the contributed sessions. It is organized in IX parts as follows:

Part I Space Discretization Methods for PDEs
Part II Finite Element Methods
Part III Discontinuous Galerkin Methods for PDEs
Part IV Numerical Linear Algebra and High Performance Computing
Part V Reduced Order Modeling
Part VI Problems with Singularities
Part VII Computational Fluid Dynamics
Part VIII Computational Methods for Multi-Physics Phenomena
Part IX Miscellaneous Topics

We would like to thank all the participants for their valuable contributions and scientific discussions during the conference and to the minisymposium organizers for helping to shape the core structure of the meeting. The members of the Scientific Committee have helped us tremendously in reviewing the contributions to this proceedings. This conference would not have been possible without all the work and guidance provided by the program committee: Franco Brezzi, Miloslav Feistauer, Roland Glowinski, Gunilla Kreiss, Yuri Kuznetsov, Pekka Neittaanmaki, Jacques Periaux, Alfio Quarteroni, Rolf Rannacher, Endre Süli, and Barbara Wohlmuth. We also thank our sponsors for their generous support: Middle East Technical University, Scientific Human Resources Development Program (ÖYP) of Ministry of Development, Turkish Academy of Sciences, Oxford University Press, and Springer Verlag. We would like to acknowledge the tireless effort of ATAK Tours; Murat Uzunca, who coordinated the edition of this Proceedings; all the staff of the Institute of Applied Mathematics for their tremendous help in organizing this conference; and our students who have helped us in many ways.

This volume is dedicated to the 60th anniversary of Middle East Technical University.

Ankara, Turkey
April 2016

Bülent Karasözen
Murat Manguoğlu
Münevver Tezer-Sezgin
Serdar Göktepe
Ömür Uğur

Contents

Part I Space Discretization Methods for PDEs

Part II Finite Element Methods

Part III Discontinuous Galerkin Methods for PDEs

Part IV Numerical Linear Algebra and High Performance Computing

Part V Reduced Order Modeling

Part VI Problems with Singularities

Part VII Computational Fluid Dynamics

Part VIII Computational Methods for Multi-physics Phenomena

Part IX Miscellaneous Topics

Part I
Space Discretization Methods for PDEs

DRBEM Solution of MHD Flow and Electric Potential in a Rectangular Pipe with a Moving Lid

Münevver Tezer-Sezgin and Canan Bozkaya

Abstract We present the dual reciprocity boundary element method (DRBEM) solution of the system of equations which model magnetohydrodynamic (MHD) flow in a pipe with moving lid at low magnetic Reynolds number. The external magnetic field acts in the pipe-axis direction generating the electric potential. The solution is obtained in terms of stream function, vorticity and electric potential in the cross-section of the pipe, and the pipe axis velocity is also computed under a constant pressure gradient. It is found that fluid flow concentrates through the upper right corner forming boundary layers with the effect of moving lid and increased magnetic field intensity. Electric field behavior is changed accordingly with the insulated and conducting portions of the pipe walls. Fluid moves in the pipe-axis direction with an increasing rate of magnitude when Hartmann number increases. The boundary only nature of DRBEM provides the solution at a low computational expense.

1 Introduction

MHD is the study of the interaction of electrically conducting fluids and electromagnetic forces. It has a widespread applications in designing cooling systems with liquid metals, MHD generators, accelerators, nuclear reactors, blood flow measurements, pumps, flow meters and etc. The most widely-known applications such as MHD flow of liquid metals are considered at low magnetic Reynolds number neglecting induced magnetic field in the fluid. The corresponding physical applications are usually MHD flows inside the pipes. When the external magnetic field applies in the pipe-axis direction, due to the interaction with the electrically conducting fluid, the electric potential is generated which can be made use of in MHD generators.

M. Tezer-Sezgin (✉) • C. Bozkaya
Department of Mathematics, Middle East Technical University, 06800, Ankara, Turkey
e-mail: munt@metu.edu.tr; bcanan@metu.edu.tr

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_1

The DRBEM is a technique that offers a great advantage to solve MHD flow equations treating all the terms (including nonlinear) other than diffusion as inhomogeneity. The studies carried by BEM and DRBEM for solving the MHD equations in pipes of several cross-sections are given in [1–6]. The externally applied magnetic field in these works is taken parallel to the cross-section plane with different orientations. Han Aydın et al. [7] and Tezer-Sezgin et al. [8] have presented also stabilized FEM and BEM-FEM solutions for MHD flow in ducts and for biomagnetic fluids, respectively. Biomagnetic fluid flow in cavities (ducts) is also studied by Tzirtzilakis [9–11] by using pressure-linked pseudotransient method on a collocated grid and finite volume method with SIMPLE algorithm, respectively.

In this paper, MHD flow in a pipe imposed to an external magnetic field in the direction of the pipe-axis is simulated using DRBEM in the cross-section of the pipe as a two-dimensional problem. The electric potential and pipe-axis velocity are also obtained with DRBEM. The boundary only nature of DRBEM gives efficient solution even by using constant elements with considerably small computational cost compared to other numerical methods.

2 The Physical Problem and Mathematical Formulation

The steady flow of an incompressible, electrically conducting, viscous fluid in a pipe in the presence of an external magnetic field acting in the pipe-axis direction is considered.

The governing dimensionless MHD equations are [12, 13]

$$\frac{1}{N}(\boldsymbol{u}.\nabla)\boldsymbol{u} - \frac{1}{M^2}\nabla^2\boldsymbol{u} + \frac{1}{N}\nabla p = \boldsymbol{B}\times\nabla\phi + \boldsymbol{B}\times(\boldsymbol{B}\times\boldsymbol{u}) \tag{1}$$

$$\nabla.\boldsymbol{u} = 0, \quad \nabla.\boldsymbol{B} = 0, \quad E = -\nabla\phi, \quad \nabla^2\phi = \mathrm{div}\,(\boldsymbol{u}\times\boldsymbol{B}) \tag{2}$$

where $\boldsymbol{u} = (u_x, u_y, u_z)$, p, $\boldsymbol{B} = (0, 0, B_0)$, ϕ are the fluid velocity, pressure, magnetic field and the electric potential, respectively. M and N are Hartmann and Stuart numbers given by $M = B_0 L\frac{\sqrt{\sigma}}{\sqrt{\rho\nu}}$, $N = \sigma B_0^2\frac{L}{\rho U_0}$ where σ, ρ, ν are the electrical conductivity, density and kinematic viscosity of the fluid, L and U_0 are the characteristic length and the velocity, and B_0 is the intensity of the applied magnetic field. Induced magnetic field is neglected due to the low magnetic Reynolds number, and $M^2 = N\,Re$, Re being fluid Reynolds number.

Flow is two-dimensional in the cross-section of the pipe (see Fig. 1) giving

$$\frac{\partial u_x}{\partial x} + \frac{\partial u_y}{\partial y} = 0 \tag{3}$$

$$\frac{1}{N}\left(u_x\frac{\partial u_x}{\partial x} + u_y\frac{\partial u_x}{\partial y}\right) - \frac{1}{M^2}\nabla^2 u_x + \frac{1}{N}\frac{\partial p}{\partial x} = -\frac{\partial\phi}{\partial y} - u_x \tag{4}$$

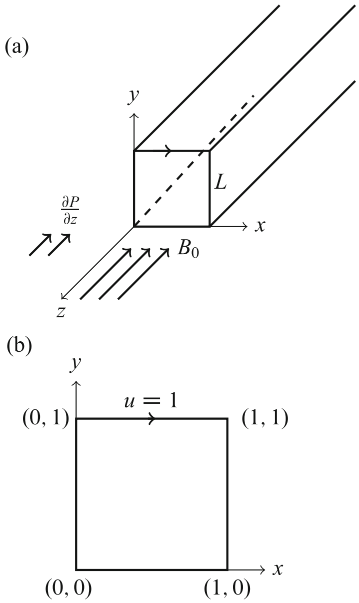


Fig. 1 (**a**) Problem domain and (**b**) cross-section of the pipe

$$\frac{1}{N}\left(u_x\frac{\partial u_y}{\partial x}+u_y\frac{\partial u_y}{\partial y}\right)-\frac{1}{M^2}\nabla^2 u_y+\frac{1}{N}\frac{\partial p}{\partial y}=\frac{\partial \phi}{\partial x}-u_y \tag{5}$$

$$\frac{1}{N}\left(u_x\frac{\partial u_z}{\partial x}+u_y\frac{\partial u_z}{\partial y}\right)-\frac{1}{M^2}\nabla^2 u_z=-\frac{1}{N}\frac{\partial P}{\partial z} \tag{6}$$

where the pressure $P = p(x, y) + P_z(z)$ is divided into the cross-section pressure $p(x, y)$ and, the pipe-axis pressure $P_z(z)$ with constant $\dfrac{\partial P_z}{\partial z}$.

Introducing stream function and vorticity in two-dimensional cross-section as

$$u_x=\frac{\partial \psi}{\partial y},\quad u_y=-\frac{\partial \psi}{\partial x},\quad w=\frac{\partial u_y}{\partial x}-\frac{\partial u_x}{\partial y}$$

we have

$$\nabla^2\psi=-w \tag{7}$$

$$\nabla^2\phi=w \tag{8}$$

$$\frac{1}{N}\left(u_x\frac{\partial w}{\partial x}+u_y\frac{\partial w}{\partial y}\right)-\frac{1}{M^2}\nabla^2 w=0 \tag{9}$$

$$\frac{1}{N}\left(u_x\frac{\partial u_z}{\partial x}+u_y\frac{\partial u_z}{\partial y}\right)-\frac{1}{M^2}\nabla^2 u_z=-\frac{1}{N}\frac{\partial P_z}{\partial z}\,. \tag{10}$$

On the boundary of the cavity, stream function is a constant due to the known velocity value, electric potential or its normal derivative is zero according to insulated or conducting portions, and the vorticity is not known.

3 DRBEM Application

DRBEM treats all the right hand side terms of Eqs. (7), (8), (9), and (10) as inhomogeneity, and an approximation for this inhomogeneous term as proposed [14] is

$$b \approx \sum_{j=1}^{K+L} \alpha_j f_j = \sum_{j=1}^{K+L} \alpha_j \nabla^2 \hat{u}_j$$

where K and L are the numbers of boundary and interior nodes, α_j are sets of initially unknown coefficients, and the f_j are approximating radial basis functions linked to particular solutions $\hat{u}_j$ with $\nabla^2 \hat{u}_j = f_j$. The radial basis functions f_j are usually chosen as polynomials of distance between the source point (x_i, y_i) and the field point (x_j, y_j) as $r_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$.

DRBEM transforms differential equations defined in a domain Ω to integral equations on the boundary $\partial\Omega$. For this, differential equation is multiplied by the fundamental solution $u^* = -\ln(r)/2\pi$ of Laplace equation and integrated over the domain. Using Divergence theorem for the Laplacian terms on both sides of the equation, domain integrals are transformed to boundary integrals.

For the discretization of the boundary, constant elements are used to obtain DRBEM matrix-vector form for Eqs. (7), (8), (9), and (10) as

$$(H\psi - G\frac{\partial\psi}{\partial n}) = (H\hat{U} - G\hat{Q})F^{-1}\{-w\} \tag{11}$$

$$(H\phi - G\frac{\partial\phi}{\partial n}) = (H\hat{U} - G\hat{Q})F^{-1}\{w\} \tag{12}$$

$$\frac{N}{M^2}\left(Hw - G\frac{\partial w}{\partial n}\right) = (H\hat{U} - G\hat{Q})F^{-1}\left\{u_x\frac{\partial w}{\partial x} + u_y\frac{\partial w}{\partial y}\right\} \tag{13}$$

$$\frac{N}{M^2}\left(Hu_z - G\frac{\partial u_z}{\partial n}\right) = (H\hat{U} - G\hat{Q})F^{-1}\left\{u_x\frac{\partial u_z}{\partial x} + u_y\frac{\partial u_z}{\partial y} + \frac{\partial P_z}{\partial z}\right\} . \tag{14}$$

Equations (11), (12), (13), and (14) are solved iteratively with an initial vorticity. With the computed ψ, the velocity components $u_x = \dfrac{\partial\psi}{\partial y}$ and $u_y = -\dfrac{\partial\psi}{\partial x}$ are computed using coordinate matrix F with entries $f_{ij} = 1 + r_{ij}$ as

$$u_x = \frac{\partial F}{\partial y} F^{-1} \psi, \quad u_y = -\frac{\partial F}{\partial x} F^{-1} \psi$$

and substituted in the vorticity and pipe-axis velocity equations. All the other space derivatives are computed using F matrix.

4 Numerical Results

The problem geometry is the lid-driven cavity $\Omega = [0, 1] \times [0, 1]$ which is the cross-section of the pipe where the top layer is moving in the positive x-direction. External magnetic field $\boldsymbol{B} = (0, 0, B_0)$ applies perpendicular to Ω and generates electric potential interacting with the electrically conducting fluid in the pipe. Fluid moves with the movement of the lid and the constant pressure gradient $\dfrac{\partial P_z}{\partial z} = -8000$ opposite to pipe-axis direction. $K = 120$ and $L = 900$ constant boundary elements and interior nodes, respectively, are taken to simulate the flow and electric potential. Solution is obtained, by using linear radial basis functions $f_{ij} = 1 + r_{ij}$ in the F matrix, for increasing values of Hartmann number M, keeping Stuart number $N = 16$ fixed. Effect of M on the pipe axis-velocity u_z is also visualized.

In Fig. 2 we present streamlines, equivorticity and equipotential lines in the case of electrically conducting pipe wall ($\phi = 0$) for Hartmann number values $M = 20$, 100, 150, 200 which correspond to Reynolds numbers $Re = 25, 625, 1406, 2500$, respectively, since $M^2 = N\,Re$. It is observed that an increase in the strength of the applied magnetic field (increase in M) causes the primary vortex of streamlines to move through the center of the cavity. Recirculations appear at the lower corners and finally at the left upper corner with further increase in M and the movement of the lid to the right. Vorticity moves away from the cavity center towards the walls indicating strong vorticity gradients. The fluid begins to rotate with a constant angular velocity and it flows creating boundary layers near the top and right walls through the upper right corner. Electric potential has the same pattern and magnitudes of streamlines since $\nabla^2\phi = w$, $\nabla^2\psi = -w$ and both ψ and ϕ are zero for this case on the cavity walls.

Figure 3 shows the increase in the magnitude of the pipe-axis velocity u_z with an increase in M when $\dfrac{\partial P_z}{\partial z} = -8000$. The damping in the magnitude of u_z is seen close to the moving lid as M increases ($M = 20, 50, 100, 150$).

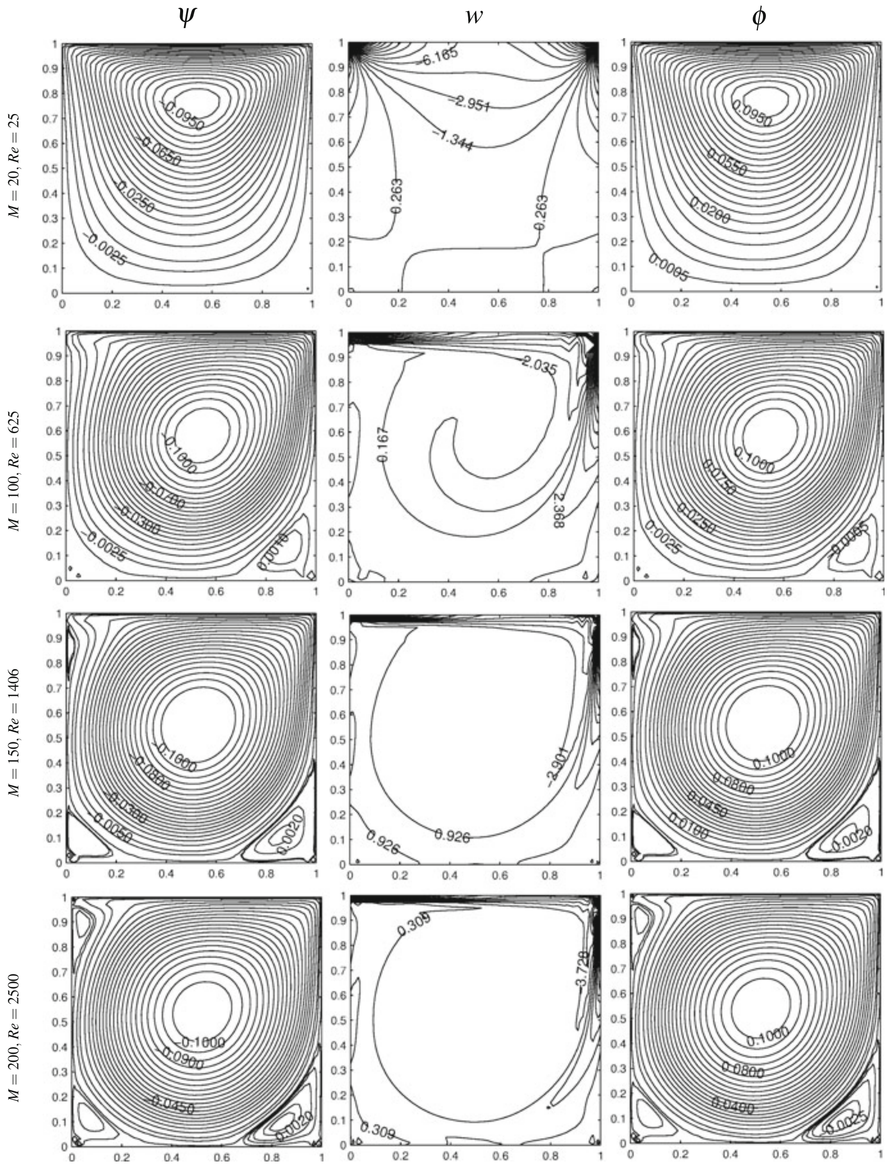
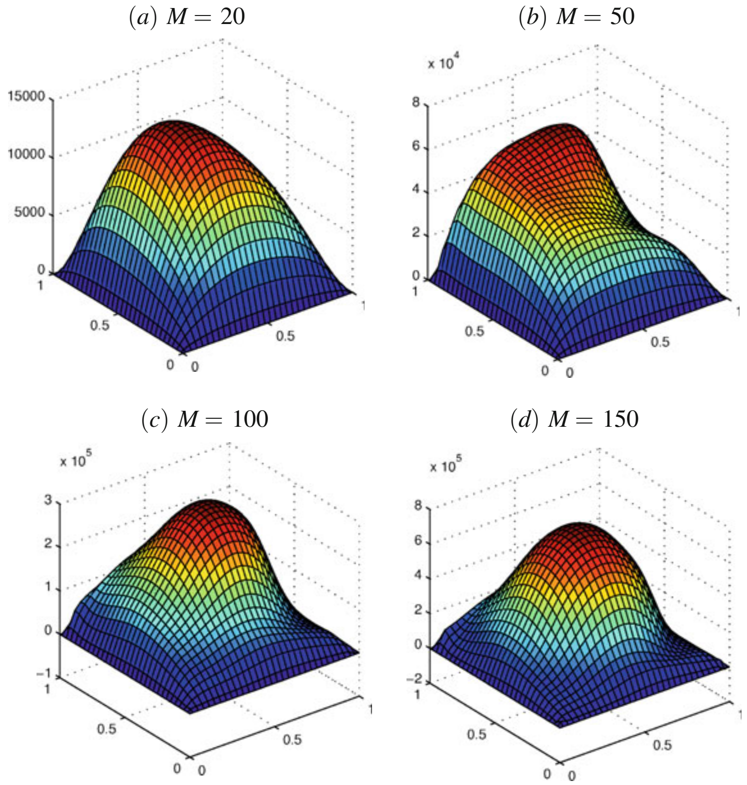


Fig. 2 Effect of Hartmann number on ψ, w and ϕ when $\phi = 0$ on the walls


$(a)\ M = 20$

$(b)\ M = 50$

$(c)\ M = 100$

$(d)\ M = 150$

Fig. 3 Pipe-axis velocity u_z when $\phi = 0$ on the walls: (**a**) $M = 20$, (**b**) $M = 50$, (**c**) $M = 100$, (**d**) $M = 150$

When the cavity walls are partly insulated and partly conducting, electric potential leaves the behavior of the flow and obeys boundary conditions on the walls for small values of M. It is seen from Fig. 4 that insulated vertical walls force the potential to touch these walls and then both the increased magnetic intensity and moving lid cause it to regain the flow behavior. On the other hand, insulated top and bottom walls give completely different pattern for the flow as traveling electric waves from the bottom to the top. Increasing Hartmann number does not change this behavior much but tends to concentrate through the upper right corner.

5 Conclusion

The MHD flow in a pipe generates electric potential when the external magnetic field applies in the pipe-axis direction. Increasing Hartmann number shows the same behavior on the flow as if increasing Reynolds number. This is the development of

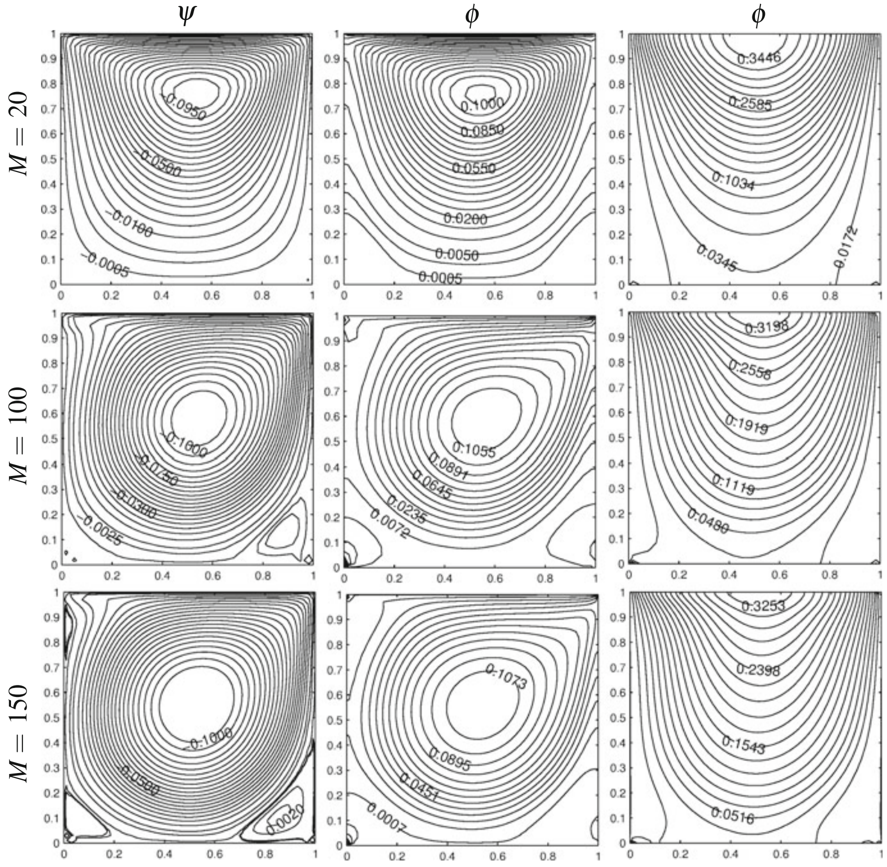


Fig. 4 Effect of Hartmann number on ψ and ϕ when $\frac{\partial \phi}{\partial n}|_{x=0,1} = 0$ (*middle*) and $\frac{\partial \phi}{\partial n}|_{y=0,1} = 0$ (*right*)

secondary flows near the lower corners and third flow close to upper right corner. This behavior is reached for much smaller *Re* values with the effect of applied magnetic field. Vorticity develops gradients on the moving lid and the right wall. Electric potential has the same behavior of the flow only when pipe walls are conducting. Pipe-axis velocity increases in magnitude with an increase in M.

References

1. C. Bozkaya, M. Tezer-Sezgin, Fundamental solution for coupled magnetohydrodynamic flow equations. J. Comput. Appl. Math. **203**, 125–144 (2007)
2. C. Bozkaya, M. Tezer-Sezgin, Boundary element method solution of magnetohydrodynamic flow in a rectangular duct with conducting walls parallel to applied magnetic field. Comput. Mech. **41**, 769–775 (2008)

3. M. Tezer-Sezgin, S. Han-Aydın, Solution of MHD flow problems using the boundary element method. Eng. Anal. Bound. Elem. **30**, 441–418 (2006)
4. M. Tezer-Sezgin, S. Han-Aydın, BEM solution of MHD flow in a pipe coupled with magnetic induction of exterior region. Computing **95**(1), 751–770 (2013)
5. M. Tezer-Sezgin, S. Han-Aydın, DRBEM solution of MHD pipe flow in a conducting medium. J. Comput. Appl. Math. **259**, 720–729 (2014)
6. C. Bozkaya, M. Tezer-Sezgin, A direct BEM solution to MHD flow in electrodynamically coupled rectangular channels. Comput. Fluids **66**, 177–182 (2012)
7. S. Han-Aydın, A.I. Neslitürk, M. Tezer-Sezgin, Two-level finite element method with a stabilizing subgrid for the incompressible MHD equations. Int. J. Numer. Methods Fluids **62**, 188–210 (2010)
8. M. Tezer-Sezgin, C. Bozkaya, Ö. Türk, BEM and FEM based numerical simulations for biomagnetic fluid. Eng. Anal. Bound. Elem. **37**(9), 1127–1135 (2013)
9. E.E. Tzirtzilakis, V.D. Sakalis, N.G. Kafoussias, P.M. Hatzikonstantinou, Biomagnetic fluid flow in a 3D rectangular duct. Int. J. Numer. Methods Fluids **44**, 1279–1298 (2004)
10. E.E. Tzirtzilakis, A mathematical model for blood flow in magnetic field. Phys. Fluids **17**, 077103 (2005)
11. E.E. Tzirtzilakis, M.A. Xenos, Biomagnetic fluid flow in a driven cavity. Meccanica **48**, 187–200 (2013)
12. U. Müller, L. Bühler, *Magnetofluiddynamics in Channels and Containers* (Springer, Berlin/New York, 2001)
13. W. Layton, H. Tran, C. Trenchea, *Numerical analysis of two partitioned methods for uncoupling evalutionary MHD flows*. Numer. Methods Partial Differ. Eq. **30**, 108301102 (2014)
14. P.W. Partridge, C.A. Brebbia, L.C. Wrobel, *The Dual Reciprocity Boundary Element Method* (Computational Mechanics Publications, Southampton/Boston, 1992)

DRBEM Solution of the Double Diffusive Convective Flow

Canan Bozkaya and Münevver Tezer-Sezgin

Abstract A numerical investigation of unsteady, two-dimensional double diffusive convection flow through a lid-driven square enclosure is carried on. The left and bottom walls of the enclosure are either uniformly or non-uniformly heated and concentrated, while the right vertical wall is maintained at a constant cold temperature. The top wall is insulated and it moves to the right with a constant velocity. The numerical solution of the coupled nonlinear differential equations is based on the use of dual reciprocity boundary element method (DRBEM) in spatial discretization and an unconditionally stable backward implicit finite difference scheme for the time integration. Due to the coupling and the nonlinearity, an iterative process is employed between the equations. The boundary only nature of the DRBEM and the use of the fundamental solution of Laplace equation make the solution process computationally easier and less expensive compared to other domain discretization methods. The study focuses on the effects of uniform and non-uniform heating and concentration of the walls for various values of physical parameters on the double-diffusive convection in terms of streamlines, isotherms and isoconcentration lines.

1 Introduction

Double-diffusive convection describes a form of convection driven by two different density gradients which have different rates of diffusion. In this sense, the double-diffusive convection generally refers to a fluid flow generated by buoyancy effects due to both temperature and solute concentration gradients. This type of flow is encountered in many engineering and geophysical applications, such as nuclear reactors, solar ponds, geothermal reservoirs, solar collectors, crystal growth in liquids, electronic cooling and chemical processing equipments. Thus, a clear understanding of the interaction between the thermal and mass or solute concentration buoyancy forces is necessary in order to control these processes.

C. Bozkaya (✉) • M. Tezer-Sezgin
Department of Mathematics, Middle East Technical University, 06800, Ankara, Turkey
e-mail: bcanan@metu.edu.tr; munt@metu.edu.tr

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_2

In the literature, the double-diffusive heat and mass transfer problems are studied mostly for square or rectangular geometries with different thermal and solute boundary conditions by using several experimental and numerical techniques. Lee et al. [1] studied experimentally the steady natural convection of salt-water solution due to horizontal temperature and concentration gradients. Cooper et al. [2] carried experiments to see the effect of buoyancy ratio R_p on the development of double-diffusive finger convection in a Hele-Shaw cell. They observed that, for low R_p fingers are rapidly developed and merge with adjacent fingers, while at higher R_p fingers are slower to evolve and do not interact as dynamically as in the lower R_p system. On the other hand, the unsteady double-diffusive convection in a square cavity was solved by Zhan et al. [3] to investigate the advantage of a hybrid method over commercial CFD codes. A finite volume approach was employed for the solution of double-diffusion flow in a cavity in [4, 5]. The effect of uniform and non-uniform heating of the walls on the double-diffusive convection in a lid-driven square cavity was analyzed by using a staggered grid finite difference method by Mahapatra et al. [6]. Alsoy et al. [7] solved the mixed convective in a lid-driven cavity and through channels with backward-facing step by the use of DRBEM.

It is seemed that, to the best of our knowledge, the double-diffusive convection in a lid-driven cavity with uniformly and non-uniformly heated and concentrated walls has not been solved by using the DRBEM which gives the solution at a considerably low computational expense due to its boundary-only nature. In the present study, we undertake this task varying the thermal Rayleigh number Ra_T and the buoyancy ratio R_p. A comprehensive study of the heat and mass transfer in terms of the flow field, temperature and concentration distribution is given in details.

2 Governing Equations

The unsteady, laminar, two-dimensional double-diffusive convection flow of an incompressible, Newtonian and viscous fluid in a lid-driven square cavity is considered. The thermo-physical properties of the fluid are assumed to be constant except the density variation in the buoyancy force, which is approximated according to the Boussinesq approximation. Thus, the non-dimensional unsteady double-diffusive convection equations in the stream function-vorticity-temperature form are written as [6]:

$$\nabla^2\psi = -\omega \tag{1}$$

$$Pr\nabla^2\omega = \frac{\partial\omega}{\partial t} + \mathbf{u}.\nabla\omega - PrRa_T\left(\frac{\partial\theta}{\partial x} + R_p\frac{\partial S}{\partial x}\right) \tag{2}$$

$$\nabla^2\theta = \frac{\partial\theta}{\partial t} + \mathbf{u}.\nabla\theta \tag{3}$$

$$\frac{1}{Le}\nabla^2 S = \frac{\partial S}{\partial t} + \mathbf{u}.\nabla S \tag{4}$$

where

$$Ra_T = \frac{g\beta_T(T_h - T_c)\ell^3}{\nu\alpha}, \quad Ra_S = \frac{g\beta_S(C_h - C_c)\ell^3}{\nu D}, \quad R_p = \frac{Ra_S}{Ra_T Le},$$

$$Pr = \frac{\nu}{\alpha}, \quad Le = \frac{\alpha}{D}.$$

Here, $\mathbf{u} = (u, v)$, ψ, w, θ, S are the velocity field, stream function, vorticity, temperature, concentration, and Pr and Le are the Prandtl number and Lewis number, respectively. The physical parameters g, α, D, ν and l given in the definitions of the thermal Rayleigh number (Ra_T) and the solutal Rayleigh number (Ra_S) are respectively the gravitational acceleration, thermal diffusivity, molecular (mass) diffusivity, kinematic viscosity and side length of the cavity. The temperatures and the concentrations at the hot and cold walls are denoted by T_h, T_c and C_h, C_c, respectively. The buoyancy ratio (R_p) is a ratio of fluid density contributions by the two solutes and defines the degree of system disequilibrium.

The corresponding dimensionless boundary conditions when $t > 0$ are shown in Fig. 1, while all unknowns are initially (at $t = 0$) taken as zero (i.e. $\psi = w = \theta = S = 0, \ 0 \le x, y \le \ell$). The thermally insulated top wall of the cavity moves to the right with a constant velocity ($\psi_y = 1$, $\psi_x = 0$), while the no-slip boundary condition is employed to the remaining walls ($\psi_x = \psi_y = 0$). Further, the bottom and left walls of the cavity are either uniformly ($\theta = S = 1$) or non-uniformly ($\theta = S = \sin\pi x$ at $y = 0$ and $\theta = S = \sin\pi y$ at $x = 0$) heated and concentrated, while the right wall is kept cold. On the other hand, the unknown boundary vorticity

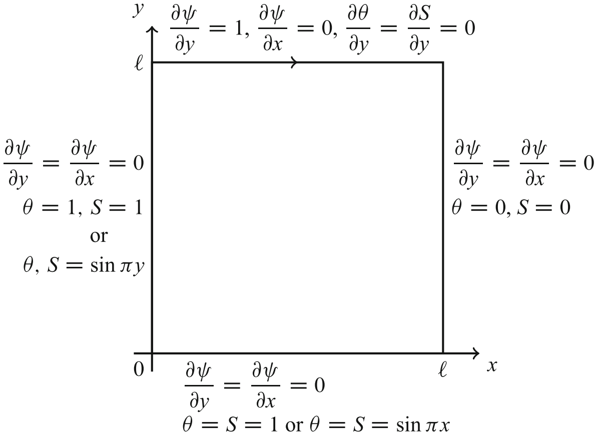

y
$\frac{\partial\psi}{\partial y} = 1, \frac{\partial\psi}{\partial x} = 0, \frac{\partial\theta}{\partial y} = \frac{\partial S}{\partial y} = 0$
ℓ
$\frac{\partial\psi}{\partial y} = \frac{\partial\psi}{\partial x} = 0$
$\theta = 1, S = 1$
or
$\theta, S = \sin\pi y$
$\frac{\partial\psi}{\partial y} = \frac{\partial\psi}{\partial x} = 0$
$\theta = 0, S = 0$
0
$\frac{\partial\psi}{\partial y} = \frac{\partial\psi}{\partial x} = 0$
$\theta = S = 1$ or $\theta = S = \sin\pi x$
ℓ
x

Fig. 1 Schematic diagram of the problem with boundary conditions

values will be obtained from the stream function equation $\Delta\psi = -w$ by using a radial basis function approximation.

3 Application of the DRBEM

The governing Eqs. (1), (2), (3), and (4) are transformed into the equivalent boundary integral equations by using DRBEM with the fundamental solution of the Laplace equation, $u^* = -\ln(r)/2\pi$, and by treating all the terms on the right hand side as inhomogeneity. An approximation for these inhomogeneous terms is

$$b \approx \sum_{j=1}^{N+L} \alpha_j f_j = \sum_{j=1}^{N+L} \alpha_j \nabla^2 \hat{u}_j$$

as proposed in [8]. Here, N and L are the numbers of boundary and interior nodes, α_j are sets of initially unknown coefficients, and f_j are approximating radial basis functions linked to particular solutions $\hat{u}_j$ with $\nabla^2\hat{u}_j = f_j$. The radial basis functions f_j are chosen as linear polynomials (i.e. $f_j = 1 + r_j$), where r_j is the distance between the source and field points.

By the use of Divergence theorem for the Laplacian terms on both sides of the equation, domain integrals are transformed into the boundary integrals. Then, constant elements are used for the discretization of the boundary, which results in the following DRBEM matrix-vector form of Eqs. (1), (2), (3), and (4)

$$H\psi - G\psi_q = C\{-\omega\}\,, \tag{5}$$

$$H\omega - G\omega_q = C\left\{\frac{1}{Pr}\left[\frac{\partial\omega}{\partial t} + \mathbf{u}.\nabla\omega - PrRa_T\left(\frac{\partial\theta}{\partial x} + R_p\frac{\partial S}{\partial x}\right)\right]\right\} \tag{6}$$

$$H\theta - G\theta_q = C\left\{\frac{\partial\theta}{\partial t} + \mathbf{u}.\nabla\theta\right\} \tag{7}$$

$$HS - GS_q = C\left\{Le\left(\frac{\partial S}{\partial t} + \mathbf{u}.\nabla S\right)\right\} \tag{8}$$

where $\psi_q = \partial\psi/\partial n$, $\omega_q = \partial\omega/\partial n$, $\theta_q = \partial\theta/\partial n$, $S_q = \partial S/\partial n$, $q^* = \partial u^*/\partial n$ and H and G are the usual DRBEM matrices. The matrix $C = (H\hat{U} - G\hat{Q})F^{-1}$ in which the matrices $\hat{U}$ and $\hat{Q}$ are constructed by taking each of the vectors $\hat{u}_j$ and $\hat{q}_j$ as columns, respectively.

The unconditionally stable backward difference integration scheme defined by

$$\frac{\partial u}{\partial t}\Big|^{n+1} = \frac{u^{n+1} - u^n}{\Delta t}$$

is used for the time integration. Here n indicates the time level. Thus, the time discretized form of DRBEM system of algebraic equations for the stream function, vorticity, temperature and concentration takes the form

$$H\psi^{n+1} - G\psi_q^{n+1} = -Cw^n \ , \tag{9}$$

$$(H - \frac{1}{Pr\Delta t}C - \frac{1}{Pr}CK)\omega^{n+1} - G\omega_q^{n+1} = -\frac{1}{Pr\Delta t}C\omega^n - Ra_T CD_x(\theta^n + R_p S^n) \tag{10}$$

$$(H - \frac{1}{\Delta t}C - CK)\theta^{n+1} - G\theta_q^{n+1} = -\frac{1}{\Delta t}C\theta^n \tag{11}$$

$$(H - \frac{Le}{\Delta t}C - LeCK)S^{n+1} - GS_q^{n+1} = -\frac{Le}{\Delta t}CS^n \tag{12}$$

where $K = u^{n+1}D_x + v^{n+1}D_y$, $D_x = \frac{\partial F}{\partial x}F^{-1}$ and $D_y = \frac{\partial F}{\partial y}F^{-1}$. The resulting system of coupled Eqs. (9), (10), (11), and (12) is solved iteratively with initial estimates of ω, θ and S. In each time level, the required space derivatives of the unknowns ψ, ω, θ and S are obtained by using coordinate matrix F as $\frac{\partial \Phi}{\partial x} = \frac{\partial F}{\partial x}F^{-1}\Phi$, $\frac{\partial \Phi}{\partial y} = \frac{\partial F}{\partial y}F^{-1}\Phi$, where Φ represents the unknowns ψ, ω, S or θ. The iterative process is terminated when a preassigned tolerance (e.g. 10^{-5}) is reached between two successive iterations.

4 Numerical Results

The unsteady double-diffusive convection in a lid-driven square cavity with uniformly and non-uniformly heated and concentrated walls is analyzed by using coupling of the DRBEM with constant elements in space with an unconditionally unstable backward difference scheme in time. The domain of problem is determined by taking the side length of the cavity $\ell = 1$. The boundary of the cavity is discretized by using maximum $N = 90$ constant boundary elements. Numerical calculations are carried out for various values of Rayleigh number ($Ra_T = 10^3, 10^5$) and buoyancy ratio ($-50 \leq R_p \leq 50$) by fixing $Pr = 0.7$ and $Le = 2$.

Figure 2 displays the effect of the Rayleigh number on the flow field, temperature and concentration at $R_p = 1$ when the bottom and left walls of the cavity are (a) uniformly (b) non-uniformly heated and concentrated. A roll with clockwise rotation is formed inside the cavity since the fluid rises up and flows down, respectively, along the hot left and cold right vertical walls. As Ra_T increases from 10^3 to 10^5, the values of stream function increase in magnitude and the flow becomes stagnant in the core of the cavity in both uniform and non-uniform cases. On the other hand, the isotherms and isoconcentration lines are dispersed in the entire cavity

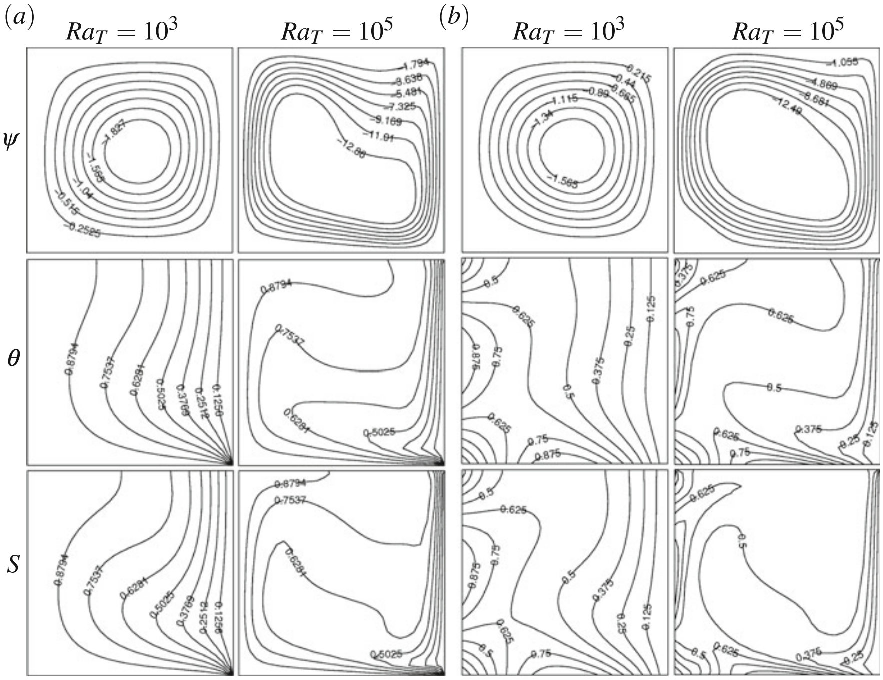


Fig. 2 Effect of the Rayleigh number Ra_T on the flow field, temperature and concentration at $R_p = 1$: *bottom* and *left walls* are (**a**) uniformly (**b**) non-uniformly heated and concentrated

at $Ra_T = 10^3$, however, lines are concentrated along the cold left vertical wall with an increase in Ra_T to 10^5 in both cases.

Effect of the buoyancy ratio R_p on the flow field, temperature and concentration at $Ra_T = 10^3$ is shown in Fig. 3 when the bottom and left walls are (a) uniformly and (b) non-uniformly heated and concentrated. In both uniform and non-uniform cases, the strength of the flow circulation decreases with a decrease in buoyancy ratio from $R_p = 50$ to $R_p = 1$ (see Fig. 2), while the stream function values increase in magnitude with a further decrease from $R_p = 1$ to $R_p = -50$. At $R_p = 50$, the contours of θ and S are mainly concentrated near the cold vertical wall and they are dispersed towards to right wall at $R_p = 1$ (see Fig. 2) in both cases. However, when $R_p = -50$, the isotherms and the isoconcentration lines are concentrated near the lower and upper half of the cold and hot vertical walls, respectively. They are almost parallel to horizontal wall in the middle part of cavity at $R_p = -50$, indicating that most of the heat transfer is carried out by heat conduction. This is due to an increase in thermal boundary layer thickness. As R_p increases boundary layer becomes thinner. This change in flow structure significantly influences the concentration field, which builds up a vertical stratification of enclosure in both uniform and non-uniform cases. The uniform heating of bottom and left walls cause a finite discontinuity for temperature distribution at one edge of bottom wall

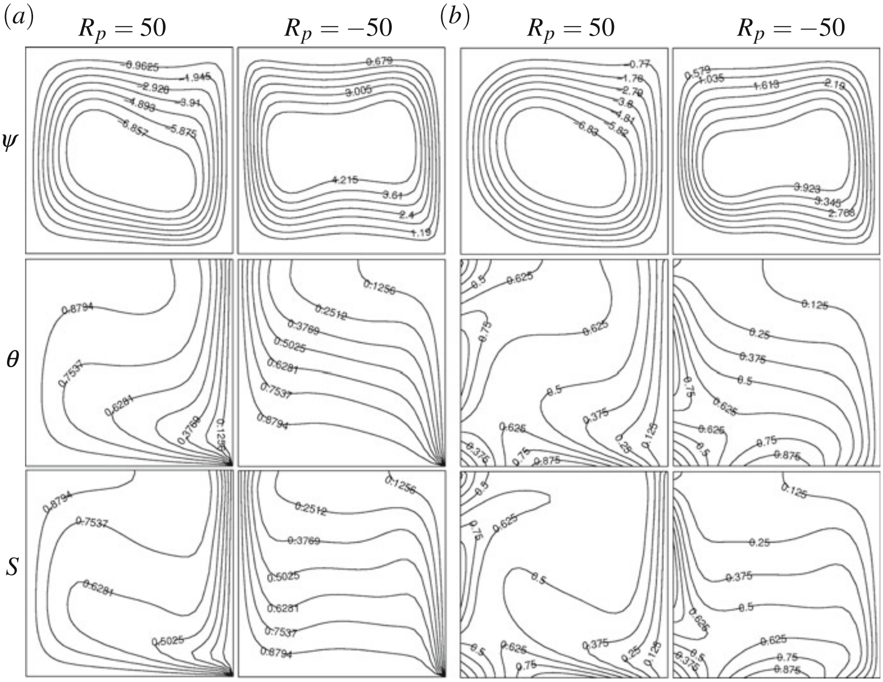


Fig. 3 Effect of the buoyancy ratio R_p on the flow field, temperature and concentration at $Ra_T = 10^3$: *bottom* and *left walls* are (**a**) uniformly (**b**) non-uniformly heated and concentrated

while the non-uniform heating removes this singularity and provides a smooth temperature distribution in the entire cavity. A similar behavior is also observed for the concentration.

The variation of the average Nusselt and Sherwood numbers at the left vertical wall ($x = 0$) and bottom wall ($y = 0$) with respect to the buoyancy ratio R_p at $Ra_T = 10^3$ is shown in Fig. 4 for (a) uniformly and (b) non-uniformly heated and concentrated walls. The average Nusselt numbers at the bottom and left walls are defined by

$$\overline{Nu}|_{y=0} = -\int_0^1 \frac{\partial \theta}{\partial y}|_{y=0}\, dx, \qquad \overline{Nu}|_{x=0} = -\int_0^1 \frac{\partial \theta}{\partial x}|_{x=0}\, dy$$

and similarly, the average Sherwood number at the bottom and left walls are

$$\overline{Sh}|_{y=0} = -\int_0^1 \frac{\partial S}{\partial y}|_{y=0}\, dx, \qquad \overline{Sh}|_{x=0} = -\int_0^1 \frac{\partial S}{\partial x}|_{x=0}\, dy\,.$$

At the bottom wall $y = 0$, the uniform and non-uniform boundary conditions produce an S-type of $\overline{Nu}$ and $\overline{Sh}$ numbers with their maximum value at right edge of

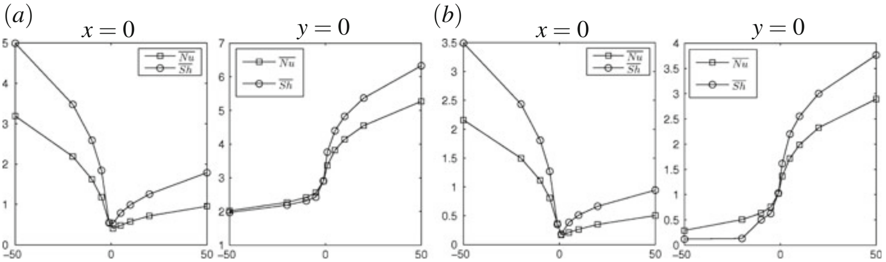


Fig. 4 The average Nusselt and Sherwood numbers at the *left vertical wall* ($x = 0$) and *bottom wall* ($y = 0$) with (**a**) uniformly and (**b**) non-uniformly heated and concentrated walls with respect to buoyancy ratio R_p at $Ra_T = 10^3$

the bottom wall. At the left vertical wall $x = 0$, $\overline{Nu}$ and $\overline{Sh}$ take the same minimum value for the uniform and non-uniform boundary conditions when $R_p = -1$.

5 Conclusion

A dual reciprocity boundary element approach in space with an implicit backward difference in time is applied for the solution of the double-diffusive convection flow in a lid-driven square cavity. The obtained results show that the flow behavior and the heat and mass transfer characteristics are significantly influenced by the use of different combination of Ra_T and R_p and they are in good agreement with the ones given in the work [6]. It is observed that the flow field is characterized by a primary circulation which moves towards the cavity walls with an increase in Ra_T. The temperature and concentration fields are significantly influenced according to the type of boundary conditions. As Rayleigh number increases and buoyancy ratio decreases, the isotherms become parallel to the adiabatic walls indicating that the heat transfer is due to the conduction. Furthermore, the heat and mass transfer rates reduce for values of $R_p < 0$, while they increase when $R_p > 0$ along the heated right wall. However, they show an S-type profile as R_p increases along the bottom wall.

References

1. J. Lee, M.T. Hyen, K.W. Kim, Natural convection in confined fluids with combined horizontal temperature and concentration gradients. Int. J. Heat Mass Tran. **31**(10), 1969–1977 (1988)
2. C.A. Cooper, R.J. Glass, S.W. Tyler, Effect of buoyancy ratio on the development of double-diffusive finger convection in a Hele-Shaw cell. Water Resour. **37**(9), 2323–2332 (2001)
3. J.M. Zhan, Y.Y. Luo, Y.S. Li, A high accuracy hybrid method for two-dimensional Navier-Stokes equations. Appl. Math. Model. **32**, 873–888 (2008)

4. M.A. Teamah, M.M. Sorour, W.M. El-Maghlany, A. Afifi, Numerical simulation of double diffusive laminar mixed convection in shallow inclined cavities with moving lid. Alex. Eng. J. **52**, 227–239 (2013)
5. B. Ghernaout, S. Bouabdallah, A. Benchatti, R. Bessaih, Effect of the buoyancy ratio on oscillatory double-diffusive convection in binary mixture. Numer. Heat Transf. Part A: Appl. Int. J. Comput. Methodol. **66**, 928–946 (2014)
6. T.R. Mahapatra, D. Pal, S. Mondal, Effects of buoyancy ratio on double-diffusive natural convection in a lid-driven cavity. Int. J. Heat Mass Transf. **57**, 771–785 (2013)
7. N. Alsoy-Akgün, M. Tezer-Sezgin, DRBEM solution of the thermo-solutal buoyancy induced mixed convection flow problems. Eng. Anal. Bound. Elem. **37**, 513–526 (2013)
8. P.W. Partridge, C.A. Brebbia, L.C. Wrobel, *The Dual Reciprocity Boundary Element Method* (Computational Mechanics Publications, Southampton/Boston, 1992)

Complete Flux Scheme for Conservation Laws Containing a Linear Source

J.H.M. ten Thije Boonkkamp, B.V. Rathish Kumar, S. Kumar, and M. Pargaei

Abstract We present an extension of the complete flux scheme for conservation laws containing a linear source. In our new scheme, we split off the linear part of the source and incorporate this term in the homogeneous flux, the remaining nonlinear part is included in the inhomogeneous flux. This approach gives rise to modified homogeneous and inhomogeneous fluxes, which reduce to the classical fluxes for vanishing linear source. On the other hand, if the linear source is large, the solution of the underlying boundary value problem is oscillatory, resulting in completely different numerical fluxes. We demonstrate the performance of the homogeneous flux approximation.

1 Introduction

Conservation laws are ubiquitous in science and engineering, describing a wide variety of phenomena, such as chemically reacting flow, electrical discharges in gases, transport in porous media etc. These conservation laws are often of advection-diffusion-reaction type, describing the interplay between different processes such as advection or drift, diffusion or conduction and (chemical) reactions or impact ionization. We restrict ourselves to stationary conservation laws.

Numerical simulation of these equations requires sophisticated space discretization methods and efficient (iterative) solvers for the resulting algebraic system. For space discretization of the conservation law we employ the finite volume method (FVM); see [2] for a detailed account. For the numerical fluxes in the discrete conservation law there exist many schemes. Basic schemes are the central

J.H.M. ten Thije Boonkkamp (✉)
Department of Mathematics and Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands
e-mail: j.h.m.tenthijeboonkkamp@tue.nl

B.V. Rathish Kumar • S. Kumar • M. Pargaei
Department of Mathematics and Statistics, Indian Institute of Technology Kanpur, Kanpur 208016, India
e-mail: bvrk@iitk.ac.in; sunilku@iitk.ac.in; meenap@iitk.ac.in

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_3

difference and upwind schemes. The central difference scheme is prone to spurious oscillations for dominant advection while the upwind discretization is too diffusive. To remedy this, exponentially fitted schemes were introduced, see [6] and the many references therein. These schemes combine the central difference and upwind schemes in such a way that the resulting discretization reproduces the exact solution of a local one-dimensional and homogeneous boundary value problem (BVP). Exponentially fitted schemes are especially useful for singularly perturbed problems [4]. Moreover, these schemes are applied in various disciplines in computational physics, such as the numerical simulation of reactive flow or plasmas. In the field of semiconductor device simulation the exponetially fitted scheme is often referred to as the Scharfetter-Gummel scheme; see e.g. [7] for a general introduction. An ingenious generalization to nonlinear convection-diffusion problems was introduced in [3], where an (iterative) procedure is proposed to compute the numerical flux from a nonlinear, but homogeneous, BVP. Another generalization is the complete flux scheme, where the flux is derived from a local BVP for the *entire* equation, including the source term [8]. Consequently, the numerical flux can be written as the superposition of a homogeneous flux, which is the exponentially fitted flux corresponding to the advection-diffusion operator, and an inhomogeneous flux, taking into account the effect of the source term.

In this contribution, we extend the derivation of the complete flux scheme to conservation laws containing a linear source. We split off the linear part and incorporate this term in the *homogeneous* flux. To that purpose, we solve the corresponding homogeneous boundary value problem, which describes the balance between advection, diffusion and a linear source. The remaining (nonlinear) part of the source is included in the inhomogeneous flux, as usual. The modified homogeneous and inhomogeneous fluxes reduce to the classical fluxes when the linear source vanishes. On the other hand, for a dominant linear source, the solution of the underlying boundary value problem exhibits oscillatory behaviour, resulting in completely different fluxes. A similar scheme is presented in [5] for the special case that the characteristic equation of the local BVP has two distinct real roots. Our scheme also allows for double real or complex (conjugate) roots.

Thus, we consider the model advection-diffusion-reaction equation

$$\frac{\mathrm{d}}{\mathrm{d}x}\Big(u\varphi - \varepsilon\frac{\mathrm{d}\varphi}{\mathrm{d}x}\Big) = c\varphi + s(\varphi), \tag{1}$$

where, for example, u is an advection velocity, $\varepsilon \geq \varepsilon_{\min} > 0$ a diffusion coefficient, $c\varphi$ the linear part of the source, and $s(\varphi)$ the remaining (nonlinear) source. The unknown φ might be the mass fraction of one of the constituent species in a reacting flow or plasma. Associated with (1) we introduce the flux f, which is defined by

$$f = u\varphi - \varepsilon\frac{\mathrm{d}\varphi}{\mathrm{d}x}. \tag{2}$$

The conservation law (1) can be concisely written as $\mathrm{d}f/\mathrm{d}x = c\varphi + s(\varphi)$ with the flux f defined in (2). In the FVM we cover the domain with a finite number of control volumes (cells) I_j of size Δx. We choose the grid points x_j, where the variable φ has to be approximated, in the cell centres. Consequently, we have $I_j := [x_{j-1/2}, x_{j+1/2}]$ with $x_{j+1/2} := \frac{1}{2}(x_j + x_{j+1})$. Integrating the equation over I_j and applying the midpoint rule for the integral of $c\varphi + s(\varphi)$, we obtain the discrete conservation law

$$F_{j+1/2} - F_{j-1/2} = \Delta x\big(c\,\varphi_j + s(\varphi_j)\big), \tag{3}$$

where $F_{j+1/2}$ and φ_j are the numerical approximation of the flux f at the cell edge $x_{j+1/2}$ and of the unknown φ at the grid point $x = x_j$, respectively. The complete flux approximation $F_{j+1/2}$ is the sum of the homogeneous flux $F^{\mathrm{h}}_{j+1/2}$ and the inhomogeneous flux $F^{\mathrm{i}}_{j+1/2}$, i.e.,

$$\begin{aligned} F_{j+1/2} &= F^{\mathrm{h}}_{j+1/2} + F^{\mathrm{i}}_{j+1/2} \\ &= \alpha_{j+1/2}\varphi_j - \beta_{j+1/2}\varphi_{j+1} + \Delta x\big(\gamma_{j+1/2}s(\varphi_j) + \delta_{j+1/2}s(\varphi_{j+1})\big). \end{aligned} \tag{4}$$

The coefficients $\alpha_{j+1/2}$ and $\beta_{j+1/2}$ depend on the homogeneous differential operator, containing the advection-diffusion operator as well as the linear source, and the coefficients $\gamma_{j+1/2}$ and $\delta_{j+1/2}$ depend on the nonlinear source $s(\varphi)$.

We have organized our paper as follows. In Sect. 2 we derive expressions for the homogeneous flux, and subsequently in Sect. 3, we outline the derivation of the inhomogeneous flux. For the latter we reformulate equation (1) and relation (2) together as a first order ODE-system. In Sect. 4 we demonstrate the performance of the homogeneous flux scheme, and finally we present conclusions in Sect. 5.

2 Modification of the Homogeneous Flux

In this section we present the extension of the homogeneous flux scheme to equation (1). We assume in the sequel of this paper that u, ε and c are constant. The expression for the homogeneous flux $F^{\mathrm{h}}_{j+1/2}$ is then derived from the following local BVP

$$\varepsilon\varphi'' - u\varphi' + c\varphi = 0, \quad x_j < x < x_{j+1}, \tag{5a}$$

$$\varphi(x_j) = \varphi_j, \quad \varphi(x_{j+1}) = \varphi_{j+1}, \tag{5b}$$

including the linear source term $c\varphi$, where the prime ($'$) denotes differentiation with respect to x. Although the source term $c\varphi$ is included, we refer to the resulting numerical flux as homogeneous, since equation (5a) is homogeneous. The inhomogeneous flux takes into account the effect of the nonlinear source $s(\varphi)$ and will be discussed in the next section.

The characteristic equation of equation (5a) reads $\varepsilon\lambda^2 - u\lambda + c = 0$ and has discriminant $D = u^2 - 4\varepsilon c$. Let us introduce the auxiliary variables

$$d = \frac{\varepsilon c}{u^2}, \quad P = \frac{u\Delta x}{\varepsilon}, \quad S = \sqrt{\frac{|c|}{\varepsilon}}\Delta x. \tag{6}$$

In (6), P is the well-known Péclet number and S is a dimensionless parameter measuring the reaction-diffusion ratio. Moreover, in the presentation that follows we encounter the following functions:

$$B(z) = \frac{z}{\mathrm{e}^z - 1}, \quad \operatorname{sinhc}(z) = \frac{\sinh(z)}{z}, \quad \operatorname{sinc}(z) = \frac{\sin(z)}{z}. \tag{7}$$

Based on the sign of the discriminant D, we can distinguish the following three cases. First, for $D > 0$, or equivalently $d < \frac{1}{4}$, the characteristic equation has two distinct real roots $\lambda = u(1 \pm r)/(2\varepsilon)$ with $r = \sqrt{1 - 4d}$. We can solve the BVP (5) and subsequently compute the numerical flux from (2). We find

$$F^{\mathrm{h}}_{j+1/2} = \frac{\varepsilon}{\Delta x}\big(C(P;r)B(-Pr)\varphi_j - C(-P;r)B(Pr)\varphi_{j+1}\big), \tag{8a}$$

$$C(P;r) = \mathrm{e}^{P(1-2r)/4}\big(\cosh\big(\tfrac{1}{4}Pr\big) + \tfrac{1}{4}P\operatorname{sinhc}\big(\tfrac{1}{4}Pr\big)\big). \tag{8b}$$

Note that the numerical flux in (8) is reminiscent of the classical homogeneous flux, and contains ‘correction factors’ $C(P;r)$ and $C(-P;r)$. Second, for $D = 0$, and hence $d = \frac{1}{4}$ and $r = 0$, the characteristic equation has the double real root $\lambda = u/(2\varepsilon)$. We find for the numerical flux

$$F^{\mathrm{h}}_{j+1/2} = \frac{\varepsilon}{\Delta x}\big(C(P)\varphi_j - C(-P)\varphi_{j+1}\big), \tag{9a}$$

$$C(P) = \mathrm{e}^{P/4}\big(1 + \tfrac{1}{4}P\big). \tag{9b}$$

Note that the numerical flux (8) reduces to (9) for $r = 0$. Finally, for $D < 0$, or equivalently $d > \frac{1}{4}$, the characteristic equation has two complex (conjugate) roots $\lambda = u(1 \pm \mathrm{i}r)/(2\varepsilon)$ with $r = \sqrt{4d - 1}$. We simply have to replace in the expressions in (8) r by ir. For the numerical flux, for example, we find

$$F^{\mathrm{h}}_{j+1/2} = \frac{\varepsilon}{\Delta x}\big(C(P;\mathrm{i}r)B(-\mathrm{i}Pr)\varphi_j - C(-P;\mathrm{i}r)B(\mathrm{i}Pr)\varphi_{j+1}\big), \tag{10}$$

so it seems as if the numerical flux is complex-valued! However, using Euler's formula, we can show that the numerical flux is real and is given by

$$F^{\mathrm{h}}_{j+1/2} = \frac{\varepsilon}{\Delta x}\big(C(P;r)\varphi_j - C(-P;r)\varphi_{j+1}\big), \tag{11a}$$

$$C(P;r) = \mathrm{e}^{P/4}\,\frac{\cos\big(\frac{1}{4}Pr\big) + \frac{1}{4}P\,\mathrm{sinc}\big(\frac{1}{4}Pr\big)}{\mathrm{sinc}\big(\frac{1}{2}Pr\big)}. \tag{11b}$$

Note that this expression is valid provided $\frac{1}{2}|P|r < \pi$.

It is interesting to investigate some limiting cases. First, for $c = 0$, we have $D > 0$, $r = 1$ and recover the well-known homogeneous numerical flux for the advection-diffusion equation given by

$$F^{\mathrm{h}}_{j+1/2} = \frac{\varepsilon}{\Delta x}\big(B(-P)\varphi_j - B(P)\varphi_{j+1}\big), \tag{12}$$

see [8]. Next, for $\varepsilon = 0$ equation (1) is an advection-reaction equation and we also have $D > 0$ and $r = 1$. The numerical flux (8) reduces to the upwind flux. Finally, for $u = 0$ equation (1) is a diffusion-reaction equation and we have $P = 0$ and $D = -4\varepsilon c$. Consequently, we have to distinguish two different cases, i.e., $c < 0$ and $c > 0$. First, for $c < 0$ it is obvious that $D > 0$ and the numerical flux (8) reduces to

$$F^{\mathrm{h}}_{j+1/2} = -\frac{\varepsilon}{\Delta x}\,\frac{\varphi_{j+1} - \varphi_j}{\mathrm{sinhc}\big(\frac{1}{2}S\big)}. \tag{13}$$

Finally, for $c > 0$ and $D < 0$ the numerical flux (11) is given by

$$F^{\mathrm{h}}_{j+1/2} = -\frac{\varepsilon}{\Delta x}\,\frac{\varphi_{j+1} - \varphi_j}{\mathrm{sinc}\big(\frac{1}{2}S\big)}, \tag{14}$$

provided $\frac{1}{2}S < \pi$. Both expressions are in fact the central difference approximation of the flux divided by the correction factor $\mathrm{sinhc}\big(\frac{1}{2}S\big)$ or $\mathrm{sinc}\big(\frac{1}{2}S\big)$. Alternatively, we could have computed these numerical fluxes directly from the BVP (5) with $u = 0$.

3 Modification of the Inhomogeneous Flux

In this section we outline the modification of the inhomogeneous flux for equation (1); a more elaborate discussion will be presented elsewhere. To derive the integral representation for the inhomogeneous flux, it is convenient to reformulate equation (1) coupled with expression (2) for the flux as the first order ODE-system

$$\boldsymbol{v}' = \boldsymbol{A}\boldsymbol{v} + \boldsymbol{b}, \quad x_j < x < x_{j+1}, \tag{15a}$$

$$\varphi(x_j) = \varphi_j, \quad \varphi(x_{j+1}) = \varphi_{j+1}, \tag{15b}$$

where $\boldsymbol{v}, \boldsymbol{A}$, and $\boldsymbol{b}$ are given by

$$\boldsymbol{v} = \begin{pmatrix} \varphi \\ f \end{pmatrix}, \quad \boldsymbol{A} = \begin{pmatrix} \frac{u}{\varepsilon} & -\frac{1}{\varepsilon} \\ c & 0 \end{pmatrix}, \quad \boldsymbol{b} = \begin{pmatrix} 0 \\ s(\varphi) \end{pmatrix}. \tag{15c}$$

This formulation is somewhat unusual, since the second component of $\boldsymbol{v}$ is the flux, for obvious reasons, instead of the derivative φ'. The fundamental matrix

$$\boldsymbol{V} = \begin{pmatrix} \varphi_1 & \varphi_2 \\ f_1 & f_2 \end{pmatrix} \tag{16}$$

corresponding to (15) satisfies the BVP

$$\boldsymbol{V}' = \boldsymbol{A}\boldsymbol{V}, \quad x_j < x < x_{j+1}, \tag{17a}$$

$$\varphi_1(x_j) = 1,\ \varphi_1(x_{j+1}) = 0, \quad \varphi_2(x_j) = 0,\ \varphi_1(x_{j+1}) = 1. \tag{17b}$$

Note that only boundary conditions for the unknown φ are specified. A straightforward derivation shows that for $D > 0$ the solutions $\varphi_1(x)$ and $\varphi_2(x)$ are given by

$$\varphi_1(x) = \mathrm{e}^{P\sigma(x)/2} \frac{\sinh\big(\frac{1}{2}Pr(1-\sigma(x))\big)}{\sinh\big(\frac{1}{2}Pr\big)}, \tag{18a}$$

$$\varphi_2(x) = \mathrm{e}^{-P(1-\sigma(x))/2} \frac{\sinh\big(\frac{1}{2}Pr\sigma(x)\big)}{\sinh\big(\frac{1}{2}Pr\big)}, \tag{18b}$$

where $\sigma(x) = (x - x_j)/\Delta x$ is the normalized coordinate on (x_j, x_{j+1}). The corresponding (homogeneous) fluxes $f_1(x)$ and $f_2(x)$ can be readily determined from (2). Similar expressions hold for $D = 0$ or $D < 0$.

Applying variation of constants, we can derive the following representation of the solution of (15), see also [1]:

$$\boldsymbol{v}(x) = \boldsymbol{V}(x)\boldsymbol{r} + \int_{x_j}^{x_{j+1}} \boldsymbol{G}(x;y)\boldsymbol{b}(y)\,\mathrm{d}y, \quad \boldsymbol{r} = \begin{pmatrix} \varphi_j \\ \varphi_{j+1} \end{pmatrix}, \tag{19}$$

with $\boldsymbol{G}(x;y)$ the Green's function given by

$$\boldsymbol{G}(x;y) = \big(\varepsilon W(\varphi_1,\varphi_2)(y)\big)^{-1} \begin{cases} \begin{pmatrix} -\varphi_1(x) f_2(y) & \varphi_1(x)\varphi_2(y) \\ -f_1(x) f_2(y) & f_1(x)\varphi_2(y) \end{pmatrix} & \text{for} \quad x_j < y \le x, \\ \begin{pmatrix} -\varphi_2(x) f_1(y) & \varphi_2(x)\varphi_1(y) \\ -f_2(x) f_1(y) & f_2(x)\varphi_1(y) \end{pmatrix} & \text{for} \quad x < y < x_{j+1}. \end{cases} \tag{20}$$

In (20), $W(\varphi_1,\varphi_2)$ is the Wronskian of φ_1 and φ_2, which for $D > 0$ is given by

$$W(\varphi_1,\varphi_2) = \begin{vmatrix} \varphi_1 & \varphi_2 \\ \varphi_1' & \varphi_2' \end{vmatrix} = \frac{1}{\Delta x} \frac{\mathrm{e}^{P(\sigma(x)-1/2)}}{\operatorname{sinhc}\big(\frac{1}{2}Pr\big)}. \tag{21}$$

Note that the relations (19), (20), and (21) define the complete solution, i.e., the unknown φ and the flux f, on the entire interval $[x_j, x_{j+1}]$. However, we are only interested in the flux at the interface $x = x_{j+1/2}$. The second component of the term $\boldsymbol{V}(x_{j+1/2})\boldsymbol{r}$ is the homogeneous flux $F^{\mathrm{h}}_{j+1/2}$ as detailed in the previous section. The inhomogeneous flux $f^{\mathrm{i}}(x_{j+1/2})$ is the second component of the inhomogeneous term in (19) evaluated at $x_{j+1/2}$ and reads

$$\begin{aligned} f^{\mathrm{i}}(x_{j+1/2}) = {} & \frac{1}{\varepsilon} F^{\mathrm{h}}_{1,j+1/2} \int_{x_j}^{x_{j+1/2}} \frac{\varphi_2(x)s(x)}{W(\varphi_1,\varphi_2)(x)}\,\mathrm{d}x \\ & + \frac{1}{\varepsilon} F^{\mathrm{h}}_{2,j+1/2} \int_{x_{j+1/2}}^{x_{j+1}} \frac{\varphi_1(x)s(x)}{W(\varphi_1,\varphi_2)(x)}\,\mathrm{d}x. \end{aligned} \tag{22}$$

The flux values $F^{\mathrm{h}}_{1,j+1/2}$ and $F^{\mathrm{h}}_{2,j+1/2}$ correspond to $f_1(x)$ and $f_2(x)$ and follow readily from the expressions (8), (9) or (11) by substituting $\varphi_j = 1,\ \varphi_{j+1} = 0$ or $\varphi_j = 0,\ \varphi_{j+1} = 1$, respectively. Applying suitable quadrature rules, we can derive expressions for the numerical inhomogeneous flux $F^{\mathrm{i}}_{j+1/2}$.

4 Numerical Example

As an example we apply the modified homogeneous flux scheme to the following model problem

$$\frac{\mathrm{d}}{\mathrm{d}x}\Big(u\varphi - \varepsilon\frac{\mathrm{d}\varphi}{\mathrm{d}x}\Big) = c\varphi, \quad 0 < x < L, \tag{23a}$$

$$\varphi(0) = \varphi_{\mathrm{L}}, \quad \varphi(L) = \varphi_{\mathrm{R}}. \tag{23b}$$

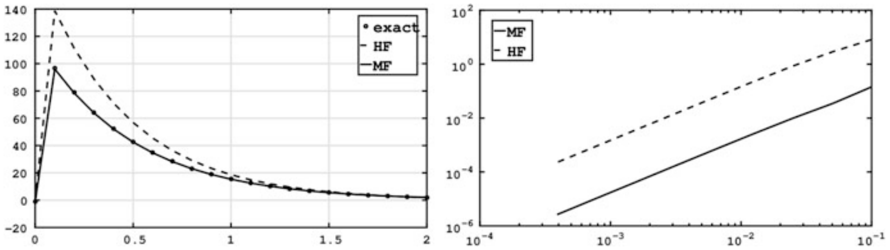


Fig. 1 Boundary layer solution: numerical solutions (*left*) and error plots (*right*). Parameter values are $u = -1$, $\varepsilon = 10^{-2}$, $c = 2$, and $\Delta x = 10^{-1}$

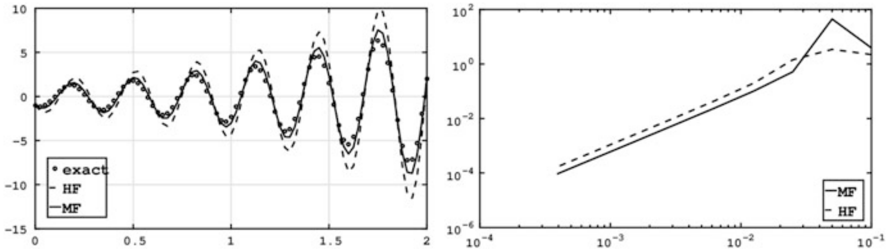


Fig. 2 Oscillatory solution: numerical solutions (*left*) and error plots (*right*). Parameter values are $u = 1$, $\varepsilon = 0.5$, and $c = 2 \times 10^2$, and $\Delta x = 2.5 \times 10^{-3}$

We consider two cases, viz. a boundary layer solution, characterized by dominant advection, and an oscillatory solution, for which the source term is dominant. For the first solution $D > 0$ and we employ the numerical flux (8), whereas for the second solution we apply the numerical flux (11) since $D < 0$.

To assess the (order) of accuracy of the modified scheme, we define the average discretization error $e(\Delta x) = \Delta x||\boldsymbol{\varphi} - \boldsymbol{\varphi}^*||_1$, with $\boldsymbol{\varphi}^*$ the exact solution of (23) restricted to the grid. A representative numerical solution and the average discretization error as function of the grid size are shown in the figures above. From the Fig. 1, we conclude that for the boundary layer solution the modified homogeneous flux scheme is much more accurate than the standard scheme, although both schemes exhibit second order convergence. On the other hand, for the oscillatory solution, the modified scheme is slightly better, as is evident from Fig. 2. Further research is needed to investigate this issue further.

5 Concluding Remarks

In this contribution we derived a new complete flux approximation scheme for conservation laws containing a linear source. We included the linear source in the homogeneous differential operator to determine the homogeneous flux. The inhomogeneous flux contains the effect of the remaining (nonlinear) part of the source.

In the derivation of the inhomogeneous flux, we reformulated the conservation law coupled with the expression for the flux as a first order ODE-system. First numerical results are encouraging, however, more testing is needed.

To be relevant for practical applications, the scheme should be extended to (at least) two-dimensional problems. This can be achieved if we include the cross-flux term as an additional source in the one-dimensional model BVP. To close the discretization, we employ the homogeneous flux scheme for the cross flux; see [8] were this idea is elaborated for the original CF scheme.

References

1. U.M. Asher, L.R. Petzold, *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations* (SIAM, Philadelphia, 1998)
2. R. Eymard, T. Gallouët, R. Herbin, Finite volume methods, in *Handbook of Numerical Analysis*, ed. by P.G. Ciarlet, J.L. Lions, vol. VII (North-Holland, Amsterdam, 2000), pp. 713–1020
3. R. Eymard, J. Fuhrmann, K. Gärtner, A finite volume scheme for nonlinear parabolic equations derived from one-dimensional local Dirichlet problems. Numer. Math. **102**, 463–495 (2006)
4. A.M. Il'in, Differencing scheme for a differential equation with a small parameter affecting the highest derivative. Math. Notes **6**, 596–602 (1969)
5. C. Luo, B.Z. Dlugogorski, B. Moghtaderi, E.M. Kennedy, Modified exponential schemes for convection-diffusion problems. Commun. Nonlinear Sci. Numer. Simul. **13**, 369–379 (2008)
6. K.W. Morton, *Numerical Solution of Convection-Diffusion Problems* (Chapman & Hall, London, 1996)
7. S. Selberherr, *Analysis and Simulation of Semiconductor Devices* (Springer, Vienna, 1984)
8. J.H.M. ten Thije Boonkkamp, M.J.H. Anthonissen, The finite volume-complete flux scheme for advection-diffusion-reaction equations. J. Sci. Comput. **46**, 47–70 (2011)

Second Order Implicit Schemes for Scalar Conservation Laws

Lisa Wagner, Jens Lang, and Oliver Kolb

Abstract The today's demands for simulation and optimization tools for water supply networks are permanently increasing. Practical computations of large water supply networks show that rather small time steps are needed to get sufficiently good approximation results – a typical disadvantage of low order methods. Having this application in mind we use higher order time discretizations to overcome this problem. Such discretizations can be achieved using so-called strong stability preserving Runge-Kutta methods which are especially designed for hyperbolic problems. We aim at approximating entropy solutions and are interested in weak solutions and variational formulations. Therefore our intention is to compare different space discretizations mostly based on variational formulations, and combine them with a second-order two-stage SDIRK method. In this paper, we will report on first numerical results considering scalar hyperbolic conservation laws.

1 Introduction

Today's demands for simulation and optimization tools for water supply networks are permanently increasing. Therefore well adapted numerical methods for the approximation of water flow in a network of pipes become important. Modeling the flow through spherical pipes, the water-hammer equations [15] or other systems of nonlinear hyperbolic balance laws can be used. Considering such systems we

L. Wagner (✉)
Department of Mathematics, TU Darmstadt, BMBF 02WER1323D EWAVE granted, Darmstadt, Germany
e-mail: wagner@mathematik.tu-darmstadt.de

J. Lang
Department of Mathematics, The Darmstadt Graduate Schools of Computational Engineering and Energy Science and Engineering, DFG TRR 154 granted, Darmstadt, Germany
e-mail: lang@mathematik.tu-darmstadt.de

O. Kolb
Department of Mathematics, University of Mannheim, Mannheim, Germany
e-mail: kolb@uni-mannheim.de

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_4

have to deal with dissipative source terms which demand implicit methods to yield fast and stable numerical methods. In this paper special emphasis is put on singly diagonally implicit Runge-Kutta (SDIRK) methods in time. In the context of hyperbolic equations, the time stepping scheme should possess the high order strong stability preserving (SSP) property [7–9]. Such methods maintain the *total variation diminishing* (TVD) property of the first order explicit Euler method. This will be described in detail in Sect. 3. Using the method of line approach, we combine a time stepping scheme with different spatial discretizations including a continuous finite element method, a finite volume method with flux limiting and a discontinuous Galerkin approach, see Sect. 2. Finally we show numerical results for linear and nonlinear test cases in Sect. 4 and compare them to the implicit box scheme developed in [12]. We end with a conclusion and an outlook to future work.

2 Different Space Discretizations

Before stating different spatial discretizations, we first introduce a prototypical scalar conservation law. Given the flux function $f : \mathbb{R} \to \mathbb{R}$ and the interval $\Omega = [0, 1]$, we aim to find $u : \Omega \times \mathbb{R}^+ \to \mathbb{R}$ solving

$$\partial_t u(x,t) + \partial_x f(u(x,t)) = 0 \quad \text{in } \Omega \times \mathbb{R}^+. \tag{1}$$

Additionally we impose the initial data $u(x, 0) = u_0(x)$ in Ω and periodic boundary conditions in space, i.e. $u(0, t) = u(1, t)$ for $t \in \mathbb{R}^+$. For analytical results see e.g. [5].

For the numerical methods we use approximate weak solutions of (1). Therefore we define $V_{per} = \{v \in L^\infty(\Omega) : v(0) = v(1) \text{ on } \partial\Omega\}$ and $W_{per} = \{w \in C^1(\Omega) : w(0) = w(1) \text{ on } \partial\Omega\}$. An integral formulation of (1) is then to find $u \in V_{per}$ such that

$$\int_\Omega \partial_t u\varphi dx + \int_\Omega \partial_x f(u)\varphi dx = 0 \quad \forall\varphi \in W_{per}(\Omega) \text{ and } t \in \mathbb{R}^+. \tag{2}$$

For given points $0 = x_0 < \ldots < x_j < \ldots < x_N = 1$, we divide Ω into intervals $I_j = [x_{j-1}, x_j]$. We denote by $h_j = x_j - x_{j-1}$ the volume of I_j. Now we can define a mesh $\Gamma_h = \{I_1, \ldots, I_N\}$ and points $x_{j-1/2} = (x_{j-1} + x_j)/2, j = 1, \ldots, N$, for later use.

2.1 *Finite Element Method*

We first discuss the finite element method although it is unusual in the context of hyperbolic problems. Nevertheless it has been reported in [11, 12] that an implicit box scheme fulfills all the desired stability and convergence properties for scalar conservation laws with possibly dissipative source terms. This scheme can be written as finite element method in space and implicit Euler method in time. Therefore our intention is to use the finite element discretization also in combination with higher order time integration. It should be noted that for continuous as well as discontinuous Galerkin methods the resolution of discontinuities which do not lie on grid points is affected by approximation errors which can result in oscillations, see Sect. 4.

For the discretization with finite elements, we approximate the solution by $u_h \in V_h \cap \mathscr{C}(\Omega) \subset V_{per}$, where V_h is the finite dimensional subspace of periodic functions consisting of first order polynomials on each I_j. We use a basis $\{\varphi_i\}_{i=0,\dots,N}$ consisting of the well-known hat functions. As test functions φ we use piecewise constants on all I_j which results in a Petrov-Galerkin method. Thus we have for all $j \in \{1, \dots, N\}$

$$\int_{I_j} \partial_t \sum_{i=0}^{N} u_i(t)\varphi_i(x)\cdot 1 \, dx + \int_{I_j} \partial_x f(\sum_{i=0}^{N} u_i(t)\varphi_i(x))\cdot 1 \, dx = 0 \tag{3}$$

where the integral can be computed exactly. Considering the periodic boundary conditions with $u_0 = u_N$, we end up with a system of ordinary differential equations for $u_j, j = 1, \dots, N$.

2.2 Finite Volume Method with Flux-Limiting

In this section we shortly introduce a finite volume scheme including a flux limiter function. In contrast to the finite element method, we compute approximations of the cell averages of the solution u. Therefore we set $K_j = [x_{j-1/2}, x_{j+1/2}]$ for $j = 1, \dots, N-1$ and the cell averages $\bar{u}_j = \frac{1}{d_j}\int_{K_j} u(x,t)dx$ with d_j as the volume of the cell K_j. For the periodic boundary conditions we define $K_N := [x_0, x_{1/2}] \cup [x_{N-1/2}, x_N]$ and set $x_{N+1/2} := x_{1/2}$. Using the conservation law (1) with initial condition above we find, after integration over K_j,

$$d_j \partial_t \bar{u}_j(t) = f(u(x_{j-1/2}, t)) - f(u(x_{j+1/2}, t))$$

on any cell K_j. Following [10, III, Section 1] and assuming $f'(u) > 0$ (which is the case for our test problems) we approximate the values of u on the cell boundaries with

$$u_{j+1/2} = \bar{u}_j + \frac{1}{2}\psi(\theta_j)(\bar{u}_{j+1} - \bar{u}_j)$$

where $\psi(\theta)$ is, e.g., the Koren limiter function

$$\psi(\theta) = \max(0, \min(2, \frac{2}{3} + \frac{1}{3}\theta, 2\theta)) \text{ with } \theta_j = \frac{\bar{u}_j - \bar{u}_{j-1}}{\bar{u}_{j+1} - \bar{u}_j}.$$

It is also possible to use different limiter functions like van Leer [10, 14]. Again repeating this procedure for every cell K_j, we end up with a system of ordinary differential equations for the cell averages $\bar{u}_j$.

2.3 Discontinuous Galerkin Method

For the discretization with discontinuous finite elements we use again the finite dimensional subspace $V_h \subset V_{per}$ consisting of first order polynomials on all I_j but without imposing the additional constraint of continuity. Consequently we have two degrees of freedom at each node, since the polynomials on each element are independent of each other. We use a basis of V_h also as test functions. Since u_h is discontinuous across the cell interfaces, we have, like for the finite volume method, to take the boundary terms from partial integration into account. Following [4] we get locally

$$\int_{I_j} \partial_t u_h \varphi_i \, dx - \int_{I_j} f(u_h) \partial_x \varphi_i \, dx + f_j(t)\varphi(x_j^-) - f_{j-1}(t)\varphi(x_{j-1}^+) = 0, \tag{4}$$

where $f_j(t) = f(u_h(x_j^-, t))$ and $f_{j-1}(t) = f(u_h(x_{j-1}^+, t))$. Considering the periodic boundary conditions we set $x_0 = x_N$. Boundary fluxes $f_j(t)$ are replaced by the numerical Roe flux $\phi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ with

$$\hat{\phi}(u_{h,j}^-, u_{h,j}^+) = \frac{f(u_{h,j}^-) + f(u_{h,j}^+)}{2} + \frac{\eta_f}{2}(u_{h,j}^- - u_{h,j}^+) \tag{5}$$

with $u_{h,j}^\pm = u_h(x_j^\pm, t)$ and

$$\eta_f = \left| \frac{f(u_{h,j}^-) - f(u_{h,j}^+)}{u_{h,j}^- - u_{h,j}^+} \right| \text{ if } u_{h,j}^- \neq u_{h,j}^+ \text{ and } f'(u_{h,j}^+) \text{ otherwise.} \tag{6}$$

Note that we recover the upwind flux in the case of the linear transport equation. This happens for any monotone flux [3]. To avoid oscillations in the nonlinear test case we limit the slope of the ansatz functions in each element with a generalized slope limiter. We decide to use the MUSCL limiter from [16] as generalized slope limiter with the *minmod* function as limiter function, see also [3, Ch. 2.4].

3 Discretization in Time

Using any of the methods above, we remain solving a system of ordinary differential equations of the form

$$\frac{\mathrm{d}}{\mathrm{d}t}U(t) = F(U(t)) \tag{7}$$

where U represents the vector of u_js and $F : \mathbb{R}^N \to \mathbb{R}^N$ results from the space discretization. Typically explicit time stepping methods are used in the context of hyperbolic conservation laws [3]. One class of higher order time discretizations methods are so called *strong-stability preserving Runge Kutta* methods (SSPRK). These were developed for solving systems resulting from hyperbolic conservation laws. We assume that there exists a constant $\Delta t_{EE} > 0$ such that the solution of the explicit Euler method applied to (7) fulfills the inequality $\|U^{n+1}\| = \|U^n + \Delta t F(U^n)\| \leq \|U^n\|$ with $U^n = U(t_n)$, for all $\Delta t \leq \Delta t_{EE}$ and a (semi)norm $\|\cdot\|$, e.g., the TV seminorm or the L_∞ norm. Problem classes (7) that fulfill this property are strong-stability preserving. Using higher order SSP methods one expects that there exists $c_{SSP} > 1$ such that the solution is SSP for all $\Delta t \leq c_{SSP}\Delta t_{EE}$. The largest coefficient c_{SSP} for which this is fulfilled is called SSP coefficient. In this paper, we use two implicit time discretization methods which are SSP. First we consider the implicit Euler method and as an example for a higher order scheme we use a two-stage singly diagonally Runge Kutta method (short: SDIRK) of order 2 with Butcher tableau [3]

$$\begin{array}{c|cc} 1/4 & 1/4 & 0 \\ 3/4 & 1/2 & 1/4 \\ \hline & 1/2 & 1/2 \end{array}.$$

Using an SDIRK scheme has the advantage that the stage values can be computed subsequently. The method we use is optimal in the sense that there exists no other two-stage SDIRK method of order two with a larger SSP coefficient. We have $c_{SSP} = 4$ [6].

4 Numerical Examples

In this chapter we test the proposed methods for the linear transport equation and for the Buckley-Leverett equation [1]. We set $\Omega = [0, 1]$, $0 \leq t \leq T = 1$ and use periodic boundary conditions in space. We use equidistant meshes and denote by $N_x = 1/\Delta x + 1$ and $N_t = T/\Delta t + 1$ the number of gridpoints in space and time, respectively. We combine the proposed SDIRK method with a finite element method (SDIRKFEM), a flux limiting method (SDIRKFLUX), a discontinuous Galerkin method (SDIRKDG), and implemented also the implicit box scheme (IBOX) from [12] for comparison. We compute for all schemes and additionally for the explicit Euler scheme applied to the different space discretizations the maximal ratio of the TV-seminorm $\|y\|_{TV} = \sum_{j=1}^{N} |\eta_j - \eta_{j-1}|$ with $y = (\eta_1, \dots, \eta_N)$, $\eta_0 = \eta_N$ for the periodic boundary conditions and for all $t \in [0, 1/8]$. We set

$$\mu(\Delta t) = \max \left\{ \frac{\|u_n\|_{TV}}{\|u_{n-1}\|_{TV}} : n \geq 1 \text{ with } n\Delta t \leq 1/8 \right\}. \tag{8}$$

We determine the corresponding stepsize for the explicit Euler method numerically to find the stepsize of the SDIRK method, see Sect. 3. Note that in case of the DG and the finite volume method we compute the ratio for the cell averages. Together with application of the slope limiter in the linear and nonlinear test case at each intermediate computation of the Runge Kutta method, it can be proven, that the SDIRKDG is TVDM (total variation diminishing in the means) [3]. All methods have been implemented using MATLAB. The systems of equations resulting from the discretization have been solved using Newton's method.

4.1 Transport Equation

In this case we have the flux function $f(u) = au$, $a = 1$ and the non-smooth box profile

$$u(x, 0) = 1 \text{ if } 1/4 \leq x \leq 3/4 \text{ and } u(x, 0) = 0 \text{ otherwise on } \Omega$$

for the initial data. The results at $t = 1$ and $N_x = 100$ are shown in Fig. 1 (left). All schemes are at least total variation bounded for the applied step sizes and all of them

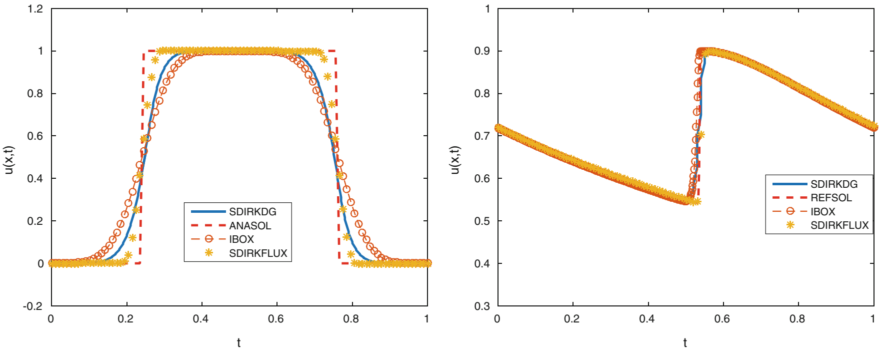


Fig. 1 *Left* (linear case): IBOX, SDIRKFLUX, SDIRKDG and analytical solution for $N_x = 100$ and $N_t = 260$ at $t = 1$. *Right* (nonlinear case): IBOX, SDIRKFLUX, SDIRKDG and reference solution for $N_x = 100$ and $N_t = 130$ and $N_x = 730$, $N_t = 65$ (for IBOX) at $t = 0.5$

Table 1 TVD property for linear case with $N_x = 100$ and $c_{SSP} = 4$

	EXPEUL		SDIRK	
	Δt	$\mu(\Delta t)$	Δt	$\mu(\Delta t)$
FEM	$2.5\cdot 10^{-6}$	1.0230	$4\cdot 2.5\cdot 10^{-6}$	1.0954
DG(1)	$2.5\cdot 10^{-6}$	1.0000	$4\cdot 2.5\cdot 10^{-6}$	1.0003
DG(1)s	$5.0\cdot 10^{-3}$	1.0000	$4\cdot 5.0\cdot 10^{-3}$	1.0000
FLUX	$2.5\cdot 10^{-3}$	1.0000	$4\cdot 2.5\cdot 10^{-3}$	1.0000

show some numerical diffusion. SDIRKFLUX is the preferably one of the three. The SDIRKFEM scheme is not shown, since it shows strong oscillations even for much finer discretizations. Nevertheless considering the IBOX scheme we obtain solutions without oscillations. Note that the IBOX scheme has to fullfill a lower bound which can be computed using [12]. For the SDIRKDG method with slope limiter there are no under- and overshoots as in the case without slope limiters. Further we compare the values of $\mu(\Delta t)$ (8) for different time stepsizes in Table 1. In the case of the SDIRKFEM TVD-stability is not achieved, but TVB-stability for a rather small time stepsize. The SDIRKDG scheme shows a difference in the size of the time stepsize between using slope limiters or not. The SDIRKFLUX scheme allows sufficiently big time stepsizes to achieve TVD-stability. Note that for the time stepsize of the explicit Euler method we have to fulfill different CFL conditions considering different spatial discretizations and test cases (for DG case see [13]).

4.2 *Buckley-Leverett Equation*

For the Buckley-Leverett equation with flux function $f(u) = 3u^2/(3u^2 + (1-u)^2)$, we consider the smooth initial profile $u(x,0) = 0.4 + 0.5\sin(\pi x)$ on Ω. The results at $t = 0.5$, $N_x = 100$ and $N_t = 130$ (for the second order schemes) are shown in Fig. 1 (right). The reference solution is computed on a very fine grid using the SDIRKFLUX scheme which again, also using the lower resolution, shows the best behaviour. The SDIRKDG method with generalized slope limiter is resolving the shock front roughly. With finer space discretization it becomes sharper. With a very small grid size ($N_x = 730$) the IBOX scheme resolves the shock very sharply. Together with $N_t = 65$ it fulfills again the lower bound mentioned above. Similar results to the linear case regarding the TVD-stability are also obtained for the nonlinear case, cf. Table 2. Table 2 shows that SDIRKDG scheme is TVD for a bigger stepsize than the SDIRKFLUX. This is not the case for nonsmooth initial data. The two test cases show that, numerically, the methods generated by the SDIRKFLUX and SDIRKDG scheme lie in the class of SSP schemes, whereas the SDIRKFEM does not. For the flux limiting method and the DG method this can be proven using Harten's lemma.

Table 2 TVD property for nonlinear case with $N_x = 100$ and $c_{SSP} = 4$

	EXPEUL		SDIRK	
	Δt	$\mu(\Delta t)$	Δt	$\mu(\Delta t)$
FEM	$2.5\cdot 10^{-7}$	1.0000	$4\cdot 2.5\cdot 10^{-6}$	1.0001
DG(1)s	$1.0\cdot 10^{-3}$	1.0000	$4\cdot 1.0\cdot 10^{-3}$	0.9991
FLUX	$1.59\cdot 10^{-4}$	1.0000	$4\cdot 1.59\cdot 10^{-4}$	1.0004

5 Conclusion

In this paper we analyzed different numerical methods for solving conservation laws obtained by the combination of a second order two-stage SDIRK method with a flux limiting method, a discontinuous Galerkin method and a finite element method. The paper includes numerical results for the linear transport equation and the Buckley-Leverett equation. We give a brief description of the used methods and focus on consistency and stability effects. Considering these properties the SDIRKFLUX method shows the best results. The SDIRKDG scheme works well with the generalized slope limiter approach, whereas the SDIRKFEM method oscillates in both cases at the discontinuities even for very small time stepsizes. For future work we plan to combine WENO (weighted essentially non-oscillatory) schemes with SDIRK or Rosenbrock methods for time integration and test our methods for solving one dimensional systems of nonlinear hyperbolic balance laws.

References

1. S.E. Buckley, M.C. Leverett, Mechanism of fluid displacements in sands. Trans. AIME **146**, 107–116 (1942)
2. J.C. Butcher, *Numerical Methods for Ordinary Differential Equations* (John Wiley & Sons, England, 2003)
3. B. Cockburn, An introduction to the discontinuous Galerkin method for convection-dominated problems, in *Advanced Numerical Approximation of Nonlinear Hyperbolic Equations*. Lecture Notes in Mathematics (Springer, Berlin/Heidelberg, 1997), pp. 151–268
4. D.A. Di Pietro, A. Ern, *Mathematical Aspects of Discontinuous Galerkin Methods*. Mathématiques et Applications, vol. 69 (Springer, Heidelberg, 2012)
5. L.C. Evans, *Partial Differential Equations* (American Mathematical Society, Rhode Island, 2010)
6. L. Ferracina, M.N. Spijker, Strong stability of singly-diagonally-implicit Runge-Kutta methods. Appl. Numer. Math. **58**, 1675–1686 (2008)
7. S. Gottlieb, C.W. Shu, Total variation diminishing Runge-Kutta schemes. Math. Comput. **67**, 73–85 (1998)
8. S. Gottlieb, C.W. Shu, E. Tadmor, Strong stability-preserving high-order time discretization methods. SIAM Rev. **43**, 89–112 (2001)
9. S. Gottlieb, D. Ketcheson, C.W. Shu, *Strong Stability Preserving Runge-Kutta and Multistep Time Discretizations* (World Scientific, Singapore, 2011)
10. W. Hundsdorfer, J.G. Verwer, *Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations*. Springer Series in Computational Mathematics, vol. 33 (Springer, Heidelberg, 2003)

11. O. Kolb, J. Lang, *Mathematical Optimization of Water Networks, Simulation and Continuous Optimization* (Birkhäuser/Springer Basel AG, Basel, 2012)
12. O. Kolb, J. Lang, P. Bales, An implict box scheme for subsonic compressible flow with dissipative source term. Numer. Algorithms **53**, 293–307 (2010)
13. E.J. Kubatko, B.A. Yeager, D.I. Ketcheson, Optimal strong-stability-preserving Runge-Kutta time discretizations for discontinuous Galerkin methods. J. Sci. Comput. **60**, 313–344 (2014)
14. R.J. LeVeque, *Numerical Methods for Conservation Laws*. Lectures in Mathematics ETH Zürich (Birkhäuser Verlag, Basel/Boston/Berlin, 1990)
15. H. Martin, R. Pohl, *Technische Hydromechanik 4* (Huss-Medien-GmbH, Berlin, 2000)
16. B. van Leer, Towards the ultimate conservation difference scheme. J. Comput. Phys. **32**, 1–136 (1974)

Flux Approximation Scheme for the Incompressible Navier-Stokes Equations Using Local Boundary Value Problems

Nikhil Kumar, J.H.M. ten Thije Boonkkamp, and Barry Koren

Abstract We present a flux approximation scheme for the incompressible Navier-Stokes equations, that is based on a flux approximation scheme for the scalar advection-diffusion-reaction equation that we developed earlier. The flux is computed from local boundary value problems (BVPs) and is expressed as a sum of a homogeneous and an inhomogeneous part. The homogeneous part depends on the balance of the convective and viscous forces and the inhomogeneous part depends on source terms included in the local BVP.

1 Introduction

The numerical solution of the incompressible Navier-Stokes equations requires appropriate spatial and temporal discretisation methods. For the spatial discretisation we consider a finite volume method (FVM), in which the conservation laws are integrated over a disjoint set of control volumes. The resulting semi-discrete conservation laws require fluxes which need to be approximated at the interfaces of the control volumes. Standard methods for the approximation of the fluxes include the *central difference* (CD) and *upwind* (UW) approximations. These methods are a consequence of two limit case solutions, i.e., the CD method results from the no-flow solution whereas the UW method corresponds to inviscid flow. This issue can be resolved if we use the exponential/hybrid scheme (as described in [1]), in which the flux approximation is based on the local balance of the convective and viscous forces, given by the solution of a homogeneous local BVP. The exponential scheme can be further extended by including the pressure gradient, the gradient of the transverse flux or the cross-flux and the time derivative of the velocity components as source terms in the local BVP. In this contribution we restrict ourselves to the steady computation of the flux and consider only the effects of including the pressure gradient and the gradient of the cross-flux.

N. Kumar (✉) • J.H.M. ten Thije Boonkkamp • B. Koren
Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands
e-mail: n.kumar@tue.nl; j.h.m.tenthijeboonkkamp@tue.nl; b.koren@tue.nl

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_5

Our objective in this paper is to formulate a flux approximation scheme, such that the computed numerical flux not only depends on the convective and viscous forces, but also includes the effects of the source terms affecting the fluid flow. Such a scheme should provide a flux approximation which is locally consistent with the corresponding conservation law. Our scheme is inspired by the complete flux-scheme for the convection-diffusion-reaction equation described in [2], in which an integral representation for the flux is derived using local BVPs for the entire equation, including the source term. In [3, 4] we have presented a similar method for the computation of the interface velocities required in the discrete convective terms using local BVPs, including the pressure gradient and the gradient of the cross-flux in the source term.

In this contribution we consider the case of two-dimensional flow. In Sect. 2 we outline the underlying FVM. Section 3 gives the integral representations of the fluxes as well as the closure of the flux scheme. In Sect. 4 we use the flux scheme to simulate flow in a lid-driven square cavity and compare the flux scheme with the CD scheme and the benchmark results. Finally, we end with a summary and concluding remarks in Sect. 5.

2 Finite Volume Method

In this section we briefly outline the FVM for the incompressible Navier-Stokes equations. Consider the two-dimensional incompressible Navier-Stokes equations

$$\nabla \cdot \boldsymbol{u} = 0, \tag{1a}$$

$$\boldsymbol{u}_t + \nabla \cdot \big(\boldsymbol{u}\boldsymbol{u} - \epsilon \nabla \boldsymbol{u}\big) = -\nabla p, \tag{1b}$$

where $\boldsymbol{u} = u\boldsymbol{e}_x + v\boldsymbol{e}_y$ is the flow velocity, p is the kinematic pressure and $\epsilon = 1/\mathrm{Re}$, with Re being the Reynolds number of the flow. For the spatial discretisation we use a staggered grid configuration as shown in Fig. 1. We have different control volumes for the discretisation of the u- and v-momentum equations. We express equation (1b) component-wise, as

$$u_t + \nabla\cdot\boldsymbol{f}^u = -\nabla \cdot (p\,\boldsymbol{e}_x), \quad \Big(\boldsymbol{f}^u := (u^2 - \epsilon u_x)\boldsymbol{e}_x + (uv - \epsilon u_y)\boldsymbol{e}_y\Big), \tag{2a}$$

$$v_t + \nabla\cdot\boldsymbol{f}^v = -\nabla \cdot (p\,\boldsymbol{e}_y), \quad \Big(\boldsymbol{f}^v := (uv - \epsilon v_x)\boldsymbol{e}_x + (v^2 - \epsilon v_y)\boldsymbol{e}_y\Big). \tag{2b}$$

Integrating equation (2a) over a control volume Ω^u and applying Gauss' theorem we get

$$\int_{\Omega^u} u_t \,\mathrm{d}A + \oint_{\partial\Omega^u} \boldsymbol{f}^u \cdot \boldsymbol{n}\,\mathrm{d}s = -\oint_{\partial\Omega^u} p\,\boldsymbol{e}_x \cdot \boldsymbol{n}\,\mathrm{d}s,$$

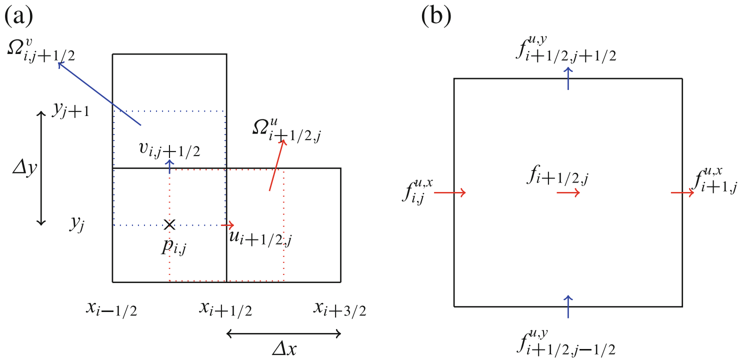


Fig. 1 (**a**) The staggered grid for the spatial discretization. (**b**) A control volume $\Omega^u_{i+1/2,j}$ for the spatial discretisation of the u-momentum equation

where $\boldsymbol{n}$ is the outward unit normal vector to the boundary $\partial\Omega^u$. This integral form of the conservation law can be approximated over the control volume $\Omega^u_{i+1/2,j}$ shown in Fig. 1 using the mid-point rule as follows:

$$\begin{aligned}\Delta y\left(f^{u,x}_{i+1,j}-f^{u,x}_{i,j}\right)+\Delta x\left(f^{u,y}_{i+1/2,j+1/2}-f^{u,y}_{i+1/2,j-1/2}\right)&=\\ -\Delta y\big(p_{i+1,j}-p_{i,j}\big)-\Delta x\,\Delta y\,(u_t)_{i+1/2,j},&\end{aligned}\tag{3}$$

where $f^{u,x} := u^2-\epsilon u_x$, $f^{u,y} := uv-\epsilon u_y$ and $f^{u,x}_{i,j} \approx f^{u,x}(x_i,y_j)$. Similarly, equation (2b) can be discretised over the control volume $\Omega^v_{i,j}$ as :

$$\begin{aligned}\Delta y\left(f^{v,x}_{i+1/2,j+1/2}-f^{v,x}_{i-1/2,j+1/2}\right)+\Delta x\left(f^{v,y}_{i,j+1}-f^{v,y}_{i,j}\right)&=\\ -\Delta x\big(p_{i,j+1}-p_{i,j}\big)-\Delta x\,\Delta y\,(v_t)_{i,j+1/2},&\end{aligned}\tag{4}$$

with $f^{v,x} := uv-\epsilon v_x$ and $f^{v,y} := v^2-\epsilon v_y$.

We begin with the approximation of the flux $f^{u,x}_{i+1,j}$ using the quasi-one-dimensional formulation of equation (2a), i.e.,

$$(f^{u,x})_x = s, \qquad \big(s := -p_x-(f^{u,y})_y-u_t\big).\tag{5}$$

Restricting the above equation to the interval $x \in [x_{i+1/2}, x_{i+3/2}]$ and $y = y_j$, the boundary conditions read

$$u(x_{i+1/2},y_j) = u_{i+1/2,j}, \quad u(x_{i+3/2},y_j) = u_{i+3/2,j}.$$

The term s acts as the source term for the flux $f^{u,x}$, giving the forces driving the flux. We have included the inertial term u_t in the source term. However, in this contribution we focus on the steady computation of the fluxes, i.e., $u_t = 0$.

The components of the fluxes $\boldsymbol{f}^u$ and $\boldsymbol{f}^v$ are nonlinear, thereby making the local BVPs nonlinear. The flux components are linearised using the interface velocities which are computed at the interface of the control volume. For example, the nonlinear term u^2 in $f^{u,x}$ is linearised as Uu, where U is the approximation of the interface velocity. The details regarding the iterative computation of the interface velocities using local BVPs are given in [3, 4]. Thus, for the approximation of the flux $f^{u,x}_{i+1,j}$ we solve the linearised local BVP

$$(Uu - \epsilon u_x)_x = -p_x - (f^{u,y})_y, \qquad x \in [x_{i+1/2}, x_{i+3/2}],\ y = y_j, \tag{6a}$$

$$u(x_{i+1/2}, y_j) = u_{i+1/2,j}, \quad u(x_{i+3/2}, y_j) = u_{i+3/2,j}. \tag{6b}$$

In the next section we give the details regarding the solution of the above local BVP.

3 Integral Representation of the Fluxes

The flux approximation scheme is based on the computation of the flux for the scalar advection-diffusion-reaction equation as described in [2]. The model equation is given by $\varphi_t + (a\varphi - \epsilon\varphi_x)_x = s$, where the scalar flux is defined as $f = a\,\varphi - \epsilon\,\varphi_x$, φ being the unknown quantity. We outline the computation of the scalar flux using a local BVP and then extend the scheme to the Navier-Stokes equations. The computation of the flux f_{i+1} at the cell edge $x_{i+1} = \frac{1}{2}(x_{i+1/2} + x_{i+3/2})$ is based on the following model BVP:

$$(a\,\varphi - \epsilon\,\varphi_x)_x = s, \quad x_{i+1/2} < x < x_{i+3/2}, \tag{7a}$$

$$\varphi(x_{i+1/2}) = \varphi_{i+1/2}, \quad \varphi(x_{i+3/2}) = \varphi_{i+3/2}. \tag{7b}$$

For the solution of the above local BVP we need the following variables:

$$\lambda := \frac{a}{\epsilon},\ P := \lambda\Delta x,\ \Lambda(x) := \int_{x_{i+1}}^{x} \lambda(\xi)\mathrm{d}\xi,\ S(x) := \int_{x_{i+1}}^{x} s(\xi)\mathrm{d}\xi,$$

with $\Delta x = x_{i+3/2} - x_{i+1/2}$ and where P is the (grid) *Péclet number*. From [2], we get that the flux f_{i+1} is the sum of a *homogeneous* (f^{h}) and an *inhomogeneous* (f^{i}) part, i.e.,

$$f_{i+1} = f^{\mathrm{h}}_{i+1} + f^{\mathrm{i}}_{i+1}, \tag{8a}$$

$$f^{\mathrm{h}}_{i+1} = \Big(e^{-\Lambda_{i+1/2}}\varphi_{i+1/2} - e^{-\Lambda_{i+3/2}}\varphi_{i+3/2}\Big) \Big/ \int_{x_{i+1/2}}^{x_{i+3/2}} \epsilon^{-1} e^{-\Lambda}\mathrm{d}x, \tag{8b}$$

$$f^{\mathrm{i}}_{i+1} = -\int_{x_{i+1/2}}^{x_{i+3/2}} \epsilon^{-1} e^{-\Lambda} S\,\mathrm{d}x \Big/ \int_{x_{i+1/2}}^{x_{i+3/2}} \epsilon^{-1} e^{-\Lambda}\mathrm{d}x. \tag{8c}$$

For the incompressible Navier-Stokes equations, we first linearise the flux component $f^{u,x}$ by defining $\tilde{f}^{u,x} = Uu - \epsilon u_x$, which can be computed using the model BVP (7). In the following we restrict ourselves to the approximation of the linearised flux $\tilde{f}^{u,x}$. Now, the source term is given by $s = -p_x - (f^{u,y})_y$. To simplify the computation of the inhomogeneous part, we make the following assumptions for the source term:

1. *Pressure gradient*: The pressure p is taken to be piecewise linear, consequently the pressure gradient is piecewise constant, given by its CD approximation:

$$p_x(x, y_j) \approx \begin{cases} (\delta_x p)_{i+1/2,j} = \frac{1}{\Delta x}(p_{i+1,j} - p_{i,j}), & x_{i+1/2} \le x \le x_{i+1}, \\ (\delta_x p)_{i+3/2,j} = \frac{1}{\Delta x}(p_{i+2,j} - p_{i+1,j}), & x_{i+1} < x \le x_{i+3/2}. \end{cases}$$

2. *Cross-flux gradient*: the gradient of the cross-flux $(f^{u,y})_y$ is taken to be piecewise constant, given by the CD approximation:

$$(f^{u,y})_y(x, y_j) \approx \begin{cases} C^u_{i+1/2,j}, & x_{i+1/2} \le x \le x_{i+1}, \\ C^u_{i+3/2,j}, & x_{i+1} < x \le x_{i+3/2}, \end{cases}$$

with

$$C^u_{i+1/2,j} = \frac{1}{\Delta y}\Big(F^{u,y}_{i+1/2,j+1/2} - F^{u,y}_{i+1/2,j-1/2}\Big),$$

$F^{u,y}$ being the numerical approximation of the linearised flux component $\tilde{f}^{u,y}$.

From the above we get that s is piecewise constant over the domain making S piecewise linear. Moreover, we also have that U and ϵ are constants on the domain $x_{i+1/2} < x < x_{i+3/2}$. Thus, evaluating expressions (8b) and (8c), we get that $F^{u,x}_{i+1,j}$, is given by

$$F^{u,x}_{i+1,j} = F^{u,x,\mathrm{h}}_{i+1,j} + F^{u,x,\mathrm{i}}_{i+1,j}, \tag{9a}$$

$$F^{u,x,\mathrm{h}}_{i+1,j} = \frac{\epsilon}{\Delta x}\big(B(-P^u)u_{i+1/2,j} - B(P^u)u_{i+3/2,j}\big), \tag{9b}$$

$$F^{u,x,\mathrm{i}}_{i+1,j} = \Delta x\big(W(-P^u)s_{i+1/2,j} - W(P^u)s_{i+3/2,j}\big), \tag{9c}$$

where $P^u = U\Delta x/\epsilon$ and

$$B(z) := \frac{z}{e^z - 1}, \quad W(z) := \frac{e^{z/2} - 1 - z/2}{z(e^z - 1)}.$$

We further split the inhomogeneous part into terms depending on the gradient of the cross-flux term ($F^{u,x,\mathrm{c}}$) and the pressure gradient ($F^{u,x,\mathrm{p}}$), using $s_{i+1/2,j} = -(\delta_x p)_{i+1/2,j} - C^u_{i+1/2,j}$ in equation (9c).

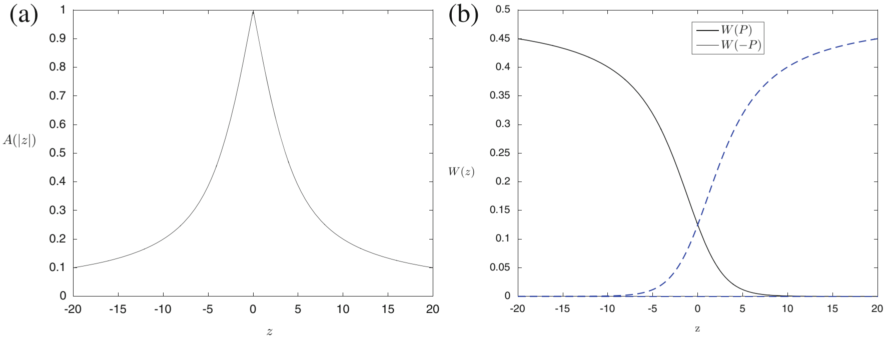


Fig. 2 Plots of the functions (**a**) $A(|z|)$, (**b**) $W(z)$ with varying Péclet numbers

Thus, we have computed the numerical flux $F^{u,x}_{i+1,j}$ as the sum of a homogeneous and an inhomogeneous part using local BVPs. Observe that the homogeneous flux component can be expressed as a weighted mean of the CD flux (F^{cd}) and the UW flux (F^{uw}) as follows

$$F^{\mathrm{h}} = \big(1 - A(|P^u|)\big)F^{\mathrm{uw}} + A(|P^u|)F^{\mathrm{cd}}, \tag{10}$$

where $A(z)$ is a weight function defined as $A(z) := 2(1 - B(z))/z$. Figure 2 for the function $A(z)$, shows that for diffusion dominated flows ($P^u \to 0$), the homogeneous scheme reduces to the CD scheme, whereas for convection dominated flows ($|P^u| \gg 1$), it reduces to the UW scheme. Analogously, the discrete source terms $s_{i+1/2,j}$ and $s_{i+3/2,j}$ involved in the inhomogeneous part have equal contributions, when the Péclet number is zero. For higher Péclet numbers the upwind source term has a larger contribution to the inhomogeneous flux part (Fig. 2).

3.1 Closure of the Scheme

So far we have derived an expression for the approximation of the flux component $F^{u,x}_{i+1,j}$. For the semi-discrete momentum equation (3) we also need to approximate the cross-flux $f^{u,y}$. For the closure of the scheme we restrict ourselves to the homogeneous flux part for the cross-flux component. Thus, the flux $f^{u,y}_{i+1/2,j+1/2}$ is computed from the local BVP :

$$(Vu - \epsilon u_y)_y = 0, \quad x = x_{i+1/2}, \ y_j \le y \le y_{j+1}, \tag{11a}$$

$$u(x_{i+1/2}, y_j) = u_{i+1/2,j}, \quad u(x_{i+1/2}, y_{j+1}) = u_{i+1/2,j+1}, \tag{11b}$$

where V is the estimate of the interface velocity at $(x_{i+1/2}, y_{j+1/2})$. On solving the above homogeneous local BVP we find that the flux is given by :

$$F^{u,y}_{i+1/2,j+1/2} = \frac{\epsilon}{\Delta y}\big(B(-P^v)u_{i+1/2,j} - B(P^v)u_{i+1/2,j+1}\big), \quad \Big(P^v = V\Delta y/\epsilon\Big).$$

Similarly, the fluxes and the cross-fluxes in equation (4) are also computed from local BVPs. The flux $f^{v,y}_{i,j+1}$ is computed using the inhomogeneous local BVP :

$$(Vv - \epsilon v_y)_y = -(\delta_y p) - (f^{v,x})_x, \quad x = x_i,\ y_{j+1/2} \le y \le y_{j+3/2}, \tag{12a}$$

$$v(x_i, y_{j+1/2}) = v_{i,j+1/2}, \quad v(x_i, y_{j+3/2}) = v_{i,j+3/2}, \tag{12b}$$

for which expressions analogous to (9) can be derived. Again, for the computation of the cross-flux $f^{v,x}_{i+1/2,j+1/2}$ we solve the homogeneous local BVP :

$$(Uv - \epsilon v_x)_x = 0, \quad x_i \le x \le x_{i+1}, y = y_{j+1/2},$$

$$v(x_i, y_{j+1/2}) = v_{i,j+1/2}, \quad v(x_{i+1}, y_{j+1/2}) = v_{i+1,j+1/2}.$$

In the following section we test the flux schemes described in this section for the lid-driven flow.

4 Numerical Results

In this section we apply the flux schemes to the flow in a lid-driven cavity, in order to assess the accuracy of the scheme. The lid-driven cavity flow is well suited to investigate the effects of including the cross-flux term in the source term. We use the results from Ghia-Ghia-Shin [5] as the reference. Figures 3 and 4 show the u-velocity profile along the vertical center-line of the cavity. In Fig. 3 we compare the homogeneous flux scheme, the 1-D flux scheme (including the pressure-gradient as the source term), and the 2-D flux scheme (including both cross-flux and the pressure-gradient), computed on a coarse 20×20 grid, with the finer grid (128×128) Ghia-Ghia-Shin solution for $\mathrm{Re} = 100$. Since the pressure gradient is practically zero across the domain, we do not see much difference between the homogeneous and the 1-D flux scheme. However the inclusion of the cross-flux term in the source term gives us a higher accuracy.

Next, we compare the flux scheme with the CD scheme for $\mathrm{Re} = 400$, (see Fig. 4). Again we compare the coarse-grid solution (20×20) with the Ghia-Ghia-Shin results. We can observe that the flux schemes exhibit higher accuracy compared to the CD scheme. The difference between the flux schemes becomes very small, with the 2-D flux scheme still being more accurate than the others though.

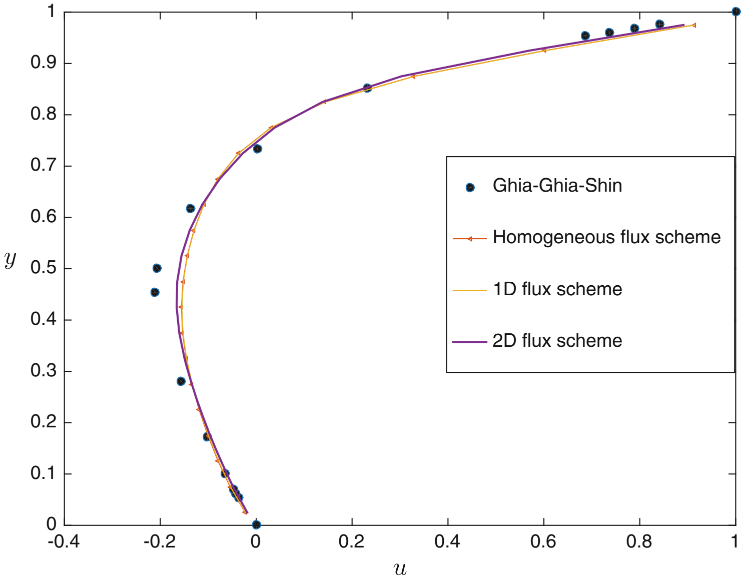


Fig. 3 u-velocity profiles along the vertical centerline of the cavity for Re $= 100$

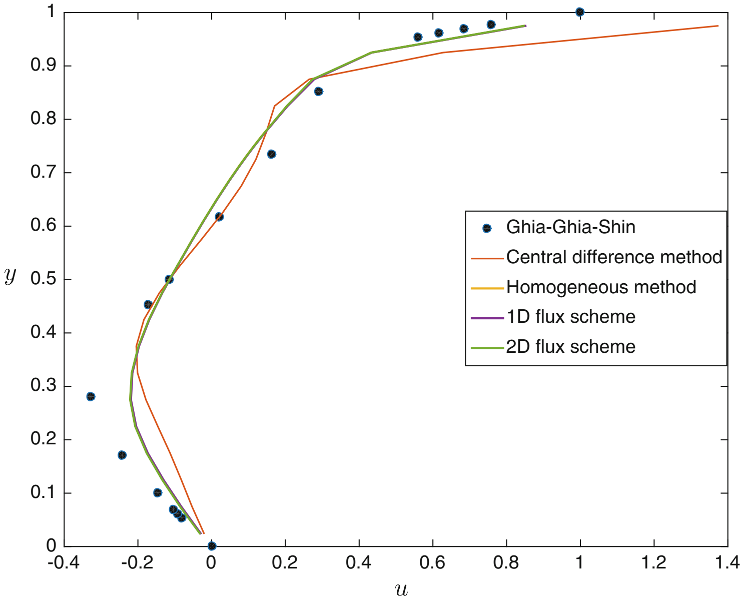


Fig. 4 u-velocity profiles along the vertical centerline of the cavity for Re $= 400$

5 Conclusion

In the preceding sections we presented methods for the approximation of the fluxes derived from local BVPs. The computed flux is the sum of a homogeneous and an inhomogeneous part. The homogeneous part is a weighted mean of the UW and CD scheme and the inhomogeneous part depends on the source term in the BVP, i.e., the pressure-gradient and the cross-flux gradient. The inclusion of the source terms provides higher accuracy to the flux approximation schemes, as observed from the case of lid-driven cavity flow.

The scheme can be further extended by including the time derivative of the velocity-component in the source term in the local BVP for the flux computation, giving the *transient flux scheme* (TFS). The TFS combined with implicit Runge-Kutta methods, should provide an accurate temporal discretisation method.

References

1. S.V. Patankar, *Numerical Heat Transfer and Fluid Flow*. Series in Computational Methods in Mechanics and Thermal Sciences (Hemisphere Publishing Corporation, New York, 1980)
2. J.H.M. ten Thije Boonkkamp, M.J.H. Anthonissen, The finite volume-complete flux scheme for advection-diffusion-reaction equations. J. Sci. Comput. **46**, 47–70 (2011)
3. N. Kumar, J.H.M. ten Thije Boonkkamp, B. Koren, A new discretization method for the convective terms in the incompressible Navier-Stokes equations, in *Finite Volumes for Complex Applications VII-Methods and Theoretical Aspects*. Springer Proceedings in Mathematics and Statistics, vol. 77 (Springer, 2014), pp. 363–371
4. N. Kumar, J.H.M. ten Thije Boonkkamp, B. Koren, A sub-cell discretization method for the convective terms in the incompressible Navier-Stokes equations, in *Spectral and High Order Methods for Partial Differential Equations ICOSAHOM2014*, Lecture Notes in Computational Science and Engineering, vol. 106 (Springer, 2014), pp. 295–303
5. U. Ghia, K.N. Ghia, C.T. Shin, High-Re solutions for incompressible flow using the Navier-Stokes equations and a multigrid method. J. Comput. Phys. **48**, 387–411 (1982)

On the Full and Global Accuracy of a Compact Third Order WENO Scheme: Part II

Oliver Kolb

Abstract Recently, we showed in (O. Kolb, *SIAM J. Numer. Anal.*, 52 (2014), pp. 2335–2355) for which parameter range the compact third order WENO reconstruction procedure introduced in (D. Levy, G. Puppo, and G. Russo, *SIAM J. Sci. Comput.*, 22 (2000), pp. 656–672) reaches the optimal order of accuracy (h^3 in the smooth case and h^2 near discontinuities). This is the case for the parameter choice $\varepsilon = Kh^q$ in the weight design with $q \leq 3$ and $pq \geq 2$, where $p \geq 1$ is the exponent used in the computation of the weights in the WENO scheme. While these theoretical results for the convergence rates of the WENO reconstruction procedure could also be validated in the numerical tests, the application within the semi-discrete central scheme of (A. Kurganov, and D. Levy, *SIAM J. Sci. Comput.*, 22 (2000), pp. 1461–1488) together with a third order TVD-Runge-Kutta scheme for the time integration did not yield a third order accurate scheme in total for $q > 2$. The aim of this follow-up paper is to explain this observation with further analytical and numerical results.

1 Introduction

We are interested in the numerical solution of hyperbolic conservation laws

$$\frac{\partial}{\partial t}u(x,t) + \frac{\partial}{\partial x}f(u(x,t)) = 0 \tag{1}$$

with given initial conditions $u(x,0) = u_0(x)$. One major difficulty arises here due to the fact that even for smooth initial data, the solutions of (the weak form of) (1) may contain discontinuities after finite time. At the same time, one is interested in resolving complex smooth solution structures with high order of accuracy. Based on the pioneering works [9, 15], the approach of so-called weighted essentially non-oscillatory (WENO) schemes allows the combination of high resolution with

O. Kolb (✉)
Department of Mathematics, University of Mannheim, A5,6 in 68131 Mannheim, Germany
e-mail: kolb@uni-mannheim.de

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_6

a stable behaviour in the presence of discontinuities. The key ingredient of such schemes is a weighting of discretization stencils or reconstruction polynomials based on smoothness indicators.

As already noted in [11], WENO reconstructions may not attain the optimal order at critical points and meanwhile there are several fixes for that problem like [2–4, 6, 8, 10, 16–18]. Based on the smoothness indicator of [11], Aràndiga et al. recently proposed in [1] to choose the parameter ε, which occurs in the denominator within the weight design, proportional to the square of the mesh size, h^2. For the compact third order WENO (CTO-WENO) reconstruction procedure introduced in [14], we recently showed in [12] that it reaches the optimal order of accuracy (h^3 in the smooth case and h^2 near discontinuities) for the parameter choice $\varepsilon = Kh^q$ with $q \leq 3$ and $pq \geq 2$, where $p \geq 1$ is the exponent used in the computation of the weights in the WENO scheme. While these theoretical results for the convergence rates of the CTO-WENO reconstruction procedure could also be validated in the numerical tests, the application within the semi-discrete central scheme of [13] together with a third order TVD-Runge-Kutta scheme from [7] for the time integration did not yield a third order accurate scheme in total for $q > 2$. Meanwhile, in [5], our results of [12] have been extended to the case of nonuniform meshes (for $\varepsilon(h) = h$ and $\varepsilon(h) = h^2$), where the dependency of ε on h is substantial. The remaining question is the explanation of the observed order reduction in the case $q > 2$ and it is the aim of this follow-up paper to explain this observation with further analytical and numerical results.

2 Numerical Scheme

We begin with a brief description of the considered discretization scheme. The underlying CTO-WENO reconstruction procedure is described for the scalar case in Sect. 2.1, and also the new results in Sect. 3 refer to the scalar case. Nevertheless, the fully discrete scheme in Sect. 2.2 is given for the system case.

2.1 Reconstruction Procedure

The CTO-WENO reconstruction procedure from [14] based on cell averages builds a core part of the analysed scheme. As in [12] we consider $u = u(x)$ as function of the spatial variable only since the procedure is independent of the time variable. Further, we assume a uniform grid with spatial grid size h, grid points $x_j = x_0 + jh$ and corresponding finite volumes $I_j = [x_j - \frac{h}{2}, x_j + \frac{h}{2}] = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$. The task is to reconstruct the function u by a piecewise polynomial approximation P given the cell averages over all I_j,

$$\bar{u}_j = \frac{1}{h} \int\limits_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(x)dx\,.$$

For this, we will use (in each cell I_j) a convex combination of three polynomials P_{L}, P_{C} and P_{R},

$$P(x) = w_{\text{L}} P_{\text{L}}(x) + w_{\text{C}} P_{\text{C}}(x) + w_{\text{R}} P_{\text{R}}(x) \tag{2}$$

with $w_i \geq 0$ for all $i \in \{\text{L}, \text{C}, \text{R}\}$ and $w_{\text{L}} + w_{\text{C}} + w_{\text{R}} = 1$. To improve the readability, we leave out the index j indicating the considered interval for the polynomials and other terms, wherever it is clear from the context.

The polynomials P_{L} and P_{R} are one-sided linear reconstructions,

$$P_{\text{L}}(x) = \bar{u}_j + \frac{\bar{u}_j - \bar{u}_{j-1}}{h}(x - x_j)\,, \qquad P_{\text{R}}(x) = \bar{u}_j + \frac{\bar{u}_{j+1} - \bar{u}_j}{h}(x - x_j)\,.$$

For the third polynomial P_{C} we need the parabola P_{opt}, which is the unique parabola that conserves the three cell averages $\bar{u}_{j-1}$, $\bar{u}_j$, $\bar{u}_{j+1}$. Then, for given (positive) constants c_{L}, c_{R} and $c_{\text{C}} = 1 - c_{\text{L}} - c_{\text{R}}$, P_{C} is chosen in such a way that

$$P_{\text{opt}}(x) = c_{\text{L}} P_{\text{L}}(x) + c_{\text{C}} P_{\text{C}}(x) + c_{\text{R}} P_{\text{R}}(x) \tag{3}$$

holds. For the weights in (2) we use

$$w_i = \frac{\alpha_i}{\sum\limits_k \alpha_k}\,, \quad \text{where} \quad \alpha_i = \frac{c_i}{(\varepsilon(h) + IS_i)^p} \quad i, k \in \{\text{L}, \text{C}, \text{R}\} \tag{4}$$

and the smoothness indicators

$$IS_i = \sum_{k=1}^{2} \int\limits_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} h^{2k-1} \big(P_i^{(k)}(x)\big)^2 dx \qquad i \in \{\text{L}, \text{C}, \text{R}\}\,. \tag{5}$$

In (4) we apply $\varepsilon(h) = Kh^q$ (with $K = 1$ in all examples) and usually $p = 2$. For the constants c_i in (3) and (4), we use $c_{\text{L}} = c_{\text{R}} = 0.25$ as in [14] and accordingly $c_{\text{C}} = 1 - c_{\text{L}} - c_{\text{R}} = 0.5$.

2.2 Fully Discrete Scheme

We now give a brief description of a complete numerical scheme to solve (1) based on the CTO-WENO reconstruction procedure presented in Sect. 2.1 (cf. [12, 13]). Note that the whole scheme can be applied to systems of conservation laws, where the reconstruction procedure can for instance be applied componentwise. First, for a given mesh size h, we average (1) over all intervals I_j. This yields the initial conditions

$$\bar{u}_j(0) = \frac{1}{h} \int\limits_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u_0(x)dx \tag{6}$$

for the cell averages in each interval I_j, and the evolution equation

$$\frac{d}{dt}\bar{u}_j(t) = -\frac{1}{h}\Big(f(u(x_{j+\frac{1}{2}}, t)) - f(u(x_{j-\frac{1}{2}}, t))\Big)\,. \tag{7}$$

Next, the fluxes $f(u(x_{j\pm\frac{1}{2}}, t))$ at the cell boundaries are replaced/approximated by a numerical flux $H_{j\pm\frac{1}{2}}(t)$ – here, corresponding to the central scheme in [13], by the local Lax-Friedrichs flux

$$H_{j+\frac{1}{2}}(t) = \frac{f(u^+_{j+\frac{1}{2}}(t)) + f(u^-_{j+\frac{1}{2}}(t))}{2} - \frac{a_{j+\frac{1}{2}}(t)}{2}\Big(u^+_{j+\frac{1}{2}}(t) - u^-_{j+\frac{1}{2}}(t)\Big) \tag{8}$$

with

$$a_{j+\frac{1}{2}}(t) = \max_{u \in C(u^-_{j+\frac{1}{2}}(t), u^+_{j+\frac{1}{2}}(t))} \rho\left(\frac{\partial f}{\partial u}(u)\right) \tag{9}$$

and

$$u^-_{j+\frac{1}{2}}(t) = P_j(x_{j+\frac{1}{2}}, t) \qquad \text{and} \qquad u^+_{j+\frac{1}{2}}(t) = P_{j+1}(x_{j+\frac{1}{2}}, t)\,.$$

The polynomials P_j and P_{j+1} are reconstructed from the cell averages at time t according to the procedure described in Sect. 2.1. Further, $\rho(A)$ denotes the spectral radius of the matrix A and $C(u^-_{j+\frac{1}{2}}(t), u^+_{j+\frac{1}{2}}(t))$ is the curve in the phase space that connects $u^-_{j+\frac{1}{2}}(t)$ and $u^+_{j+\frac{1}{2}}(t)$ via a Riemann fan.

Finally, the third order TVD Runge-Kutta scheme of [7] is used for the time integration of the semi-discretized problem

$$\frac{d}{dt}\bar{u}_j^h(t) = -\frac{1}{h}\big(H_{j+\frac{1}{2}}(t) - H_{j-\frac{1}{2}}(t)\big)$$

with approximate solution $\bar{u}^h$ and initial conditions from (6).

3 New Results

3.1 A Sufficient Condition in the Linear Case

A usual argumentation for an mth order scheme (with respect to the spatial semi-discretization) goes as follows: The exact evolution of the cell averages in each interval I_j is given by (7). Now assume that the numerical flux satisfies

$$H_{j+\frac{1}{2}}(t) = f(u(x_{j+\frac{1}{2}},t)) + d(x_{j+\frac{1}{2}},t)\,h^m + \mathcal{O}(h^{m+1}) \tag{10}$$

with a Lipschitz continuous function $d(x,t)$ with Lipschitz constant L_d (with respect to x). Then,

$$\begin{aligned}\frac{H_{j+\frac{1}{2}}(t) - H_{j-\frac{1}{2}}(t)}{h} = {}& \frac{f(u(x_{j+\frac{1}{2}},t)) - f(u(x_{j-\frac{1}{2}},t))}{h}\\ &+ \underbrace{\big(d(x_{j+\frac{1}{2}},t) - d(x_{j-\frac{1}{2}},t)\big)}_{\|\ldots\| \le L_d h} h^{m-1} + \mathcal{O}(h^m)\end{aligned}$$

and further (as desired)

$$\frac{d}{dt}\bar{u}_j(t) = -\frac{1}{h}\big(H_{j+\frac{1}{2}}(t) - H_{j-\frac{1}{2}}(t)\big) + \mathcal{O}(h^m)\,.$$

In the simplest case of a linear flux function, $f(u) = au$ with $a > 0$ (w.l.o.g.), the local Lax-Friedrichs flux (8) reduces to $H_{j+\frac{1}{2}}(t) = au^-_{j+\frac{1}{2}}(t)$ and the "sufficient condition" (10) directly reduces to an accuracy condition

$$u^-_{j+\frac{1}{2}}(t) - u(x_{j+\frac{1}{2}},t) = \tilde{d}(x_{j+\frac{1}{2}},t)\,h^m + \mathcal{O}(h^{m+1}) \tag{11}$$

with a Lipschitz continuous function $\tilde{d}(x,t) = d(x,t)/a$ (with respect to x). Further, since $u^-_{j+\frac{1}{2}}(t) = P_j(x_{j+\frac{1}{2}},t)$, we have to take a closer look at the accuracy of the reconstruction polynomials P_j given by (2). Actually, we are interested in the case

$m = 3$ (a third order scheme in the smooth case). From the proof of Theorem 2.1 in [12], we see that the deviation

$$c_i - w_i = \mathcal{O}(h)$$

from the optimal weights is essential to get third order accuracy in the reconstruction since this deviation is multiplied with the interpolation error of the single polynomials P_i, which is at least $\mathcal{O}(h^2)$ (in the smooth case). To also fulfill (11), it would be sufficient to have

$$c_i - w_i = d_i(x_j, t)\, h + \mathcal{O}(h^2) \tag{12}$$

with Lipschitz continuous functions d_i.

3.2 Failure of $q > 2$: Numerical Evidence

According to the sufficient condition (12), it should be revealing to look at $\frac{c_i - w_i}{h}$ for $h \to 0$. Therefore, we consider the initial conditions of the first "failing" example of [12] (originally from [10]),

$$u_0(x) = \sin\big(\pi x - \sin(\pi x)/\pi\big)$$

on the computational domain $x \in [-1, 1]$ (with periodic boundary conditions). For $Nx = 2^n$ grid cells with $n \in \{10, 15, 20\}$, corresponding to $h = 2 \cdot 2^{-n}$, we apply the CTO-WENO reconstruction with $\varepsilon(h) = h^3$ and evaluate $\frac{c_i - w_i}{h}$ for each cell and $i \in \{\text{L}, \text{C}, \text{R}\}$.

Figure 1 shows the corresponding results. Note the different scales on the y-axes and the different behaviour for $i \in \{\text{L}, \text{R}\}$ in comparison to $i = \text{C}$. Obviously, the quotient $\frac{c_i - w_i}{h}$ seems to be unbounded at least for $i \in \{\text{L}, \text{R}\}$ and $h \to 0$ close to the zeros of u_0' (at approximately ± 0.597). At the first view, this observation even seems to be contradictory to the results of [12], but the quotient is bounded for each fixed position x_j so that finally $w_i = c_i + \mathcal{O}(h)$ holds for arbitrary x also in the considered case $\varepsilon(h) = h^3$. Nevertheless, this behaviour is much different from the results one observes in the case $\varepsilon(h) = h^q$ with $q \leq 2$ and it obstructs the "sufficient condition" (12).

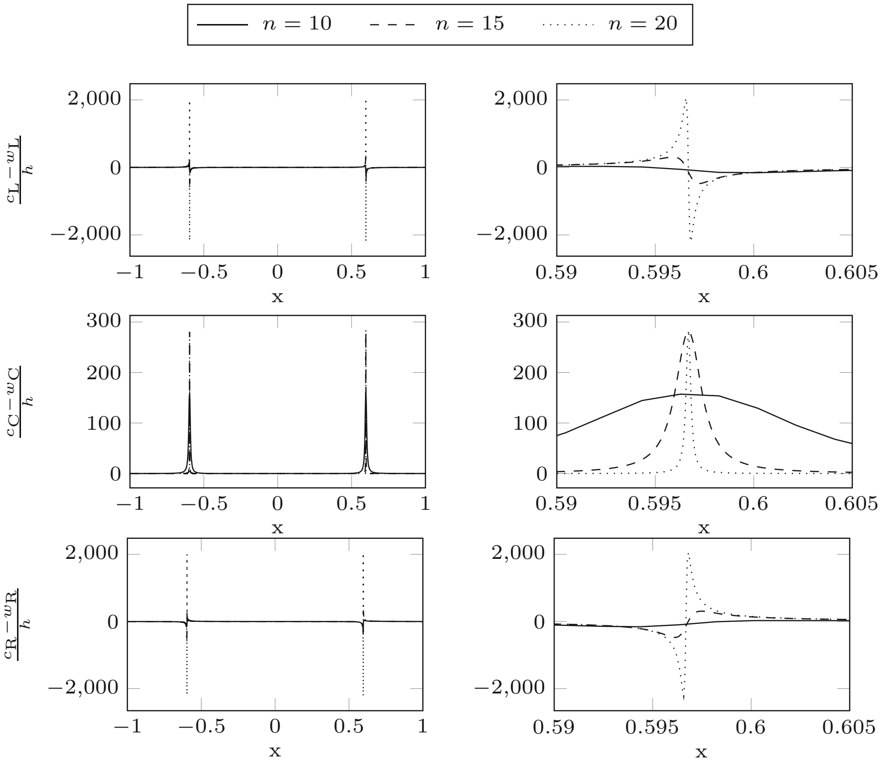


Fig. 1 Differences $\frac{c_i - w_i}{h}$ for $i \in \{\text{L}, \text{C}, \text{R}\}$, $h = 2 \cdot 2^{-n}$ with $n \in \{10, 15, 20\}$, $\varepsilon(h) = h^3$. The plots in the *right column* are zoomed from the plots in the *left column*

3.3 Failure of $q > 2$: Analytical Evidence

Next, we aim to explain the observation above from the analytical point of view. First, the proof of Theorem 3.3 in [12] gives us for any fixed x_j

$$c_i - w_i = -pf_i h^r + \mathcal{O}(h^{r+1})$$

with $r \geq 1$ and $f_i = -\sum_k c_k e_{ik}$, where the e_{ik} are given by

$$\frac{IS_i - IS_k}{\varepsilon(h) + IS_k} = e_{ik} h^r + \mathcal{O}(h^{r+1}) \, .$$

The term on the left-hand side can be expressed as

$$\frac{IS_i - IS_k}{\varepsilon(h) + IS_k} = \frac{c_{ik}h^{2s+2} + d_{ik}h^{2s+3} + \mathscr{O}(h^{2s+4})}{Kh^q + a_k h^{2s+2} + b_k h^{2s+3} + \mathscr{O}(h^{2s+4})} \tag{13}$$

with $s = s_j$ (multiplicity of the zero of u' at x_j, or 0 if $u'(x_j) \neq 0$) and appropriate constants a_i and b_i, $c_{ik} = a_i - a_k$ and $d_{ik} = b_i - b_k$ from

$$IS_i = a_i h^{2s+2} + b_i h^{2s+3} + \mathscr{O}(h^{2s+4})$$

and

$$IS_i - IS_k = c_{ik}h^{2s+2} + d_{ik}h^{2s+3} + \mathscr{O}(h^{2s+4})\,.$$

Motivated by the numerical results above, we take a closer look at the zeros of u'. Considering $s_j > 0$ for $q \in (2, 3]$ gives

$$\frac{IS_i - IS_k}{\varepsilon(h) + IS_k} = \frac{d_{ik}h^{2s+3-q} + \mathscr{O}(h^{2s+4-q})}{K + a_k h^{2s+2-q} + \mathscr{O}(h^{2s+3-q})} = \frac{d_{ik}}{K}h^{2s+3-q} + \mathscr{O}(h^{2s+4-q})\,,$$

for even s_j (where $c_{ik} = 0$ according to [12]), and for odd s_j

$$\frac{IS_i - IS_k}{\varepsilon(h) + IS_k} = \frac{c_{ik}h^{2s+2-q} + \mathscr{O}(h^{2s+3-q})}{K + a_k h^{2s+2-q} + \mathscr{O}(h^{2s+3-q})} = \frac{c_{ik}}{K}h^{2s+2-q} + \mathscr{O}(h^{2s+3-q})\,.$$

Due to the dominant role of the constant K in the denominator (and $2s+2-q \geq 1$), this case seems to be uncritical. The real problem are the points close to the zeros of u': In the case $s_j = 0$, we get for $q \in (2, 3]$

$$\frac{IS_i - IS_k}{\varepsilon(h) + IS_k} = \frac{d_{ik}h + \mathscr{O}(h^2)}{Kh^{q-2} + a_k + b_k h + \mathscr{O}(h^2)} = \frac{d_{ik}}{a_k}h + \mathscr{O}(h^{q-1})\,, \tag{14}$$

where again $c_{ik} = 0$ according to [12]. From the proof of Lemma 3.1 in [12], we know that $a_k = \left(u'(x_j)\right)^2$ here, whereas b_k and therewith d_{ik} are proportional to $u'(x_j)u''(x_j)$. For any fixed x_j with $u'(x_j) \neq 0$, Eq. (14) is sufficient to finally get third order accuracy for the CTO-WENO reconstruction, but obviously the factor $e_{ik} = \frac{d_{ik}}{a_k}$, which is proportional to $u''(x_j)/u'(x_j)$, is not bounded uniformly in x close to zeros of u' (unless also u'' vanishes in that point). This clearly explains the increase of the quotient $\frac{c_i - w_i}{h}$ for $h \to 0$ close to the zeros of u' and therewith finally leads to the observed order reduction.

Remark 1 Obviously, for $q \leq 2$ the term $\varepsilon(h) = Kh^q$ is always part of the dominant term in the denominator of (13) (as already noted in [12]) and therefore the quotient $\frac{c_i - w_i}{h}$ stays bounded in that case and even the sufficient condition (12) is fulfilled.

Remark 2 Reconsidering the scalar accuracy tests from [12], one actually observes that third order accuracy is achieved by the fully discrete scheme with $\varepsilon(h) = h^3$ apart from critical points.

4 Conclusion

The aim of this work was to explain the order reduction one observes for a fully discrete scheme based on the CTO-WENO reconstruction procedure with $\varepsilon(h) = Kh^q$ with $q \in (2, 3]$, whereas the pure spatial reconstruction is (pointwise) third order accurate. Therefore, we took a closer look at the error expansions and found numerical as well as analytical evidence for the "failure" of this parameter range. Consequently, at least for the usual choice $p = 2$ in the weight design, the region of practical interest is $q \in [1, 2]$, for which meanwhile third order accuracy has also been shown for the case of nonuniform meshes in [5].

References

1. F. Aràndiga, A. Baeza, A.M. Belda, P. Mulet, Analysis of WENO schemes for full and global accuracy. SIAM J. Numer. Anal. **49**(2), 893–915 (2011)
2. R. Borges, M. Carmona, B. Costa, W.S. Don, An improved weighted essentially non-oscillatory scheme for hyperbolic conservation laws. J. Comput. Phys. **227**(6), 3191–3211 (2008)
3. S. Bryson, D. Levy, Mapped WENO and weighted power ENO reconstructions in semi-discrete central schemes for Hamilton-Jacobi equations. Appl. Numer. Math. **56**(9), 1211–1224 (2006)
4. M. Castro, B. Costa, W.S. Don, High order weighted essentially non-oscillatory WENO-Z schemes for hyperbolic conservation laws. J. Comput. Phys. **230**(5), 1766–1792 (2011)
5. I. Cravero, M. Semplice, On the accuracy of WENO and CWENO reconstructions of third order on nonuniform meshes. J. Sci. Comput. **67**(3), 1219–1246 (2016)
6. H. Feng, F. Hu, R. Wang, A new mapped weighted essentially non-oscillatory scheme. J. Sci. Comput. **51**(2), 449–473 (2012)
7. S. Gottlieb, C.-W. Shu, Total variation diminishing Runge-Kutta schemes. Math. Comput. **67**, 73–85 (1998)
8. Y. Ha, C.H. Kim, Y.J. Lee, J. Yoon, An improved weighted essentially non-oscillatory scheme with a new smoothness indicator. J. Comput. Phys. **232**(1), 68–86 (2013)
9. A. Harten, B. Engquist, S. Osher, S.R. Chakravarthy, Uniformly high order accurate essentially non-oscillatory schemes, III. J. Comput. Phys. **71**(1), 231–303 (1987)
10. A.K. Henrick, T.D. Aslam, J.M. Powers, Mapped weighted essentially non-oscillatory schemes: achieving optimal order near critical points. J. Comput. Phys. **207**(2), 542–567 (2005)
11. G.-S. Jiang, C.-W. Shu, Efficient implementation of weighted ENO schemes. J. Comput. Phys. **126**(1), 202–228 (1996)
12. O. Kolb, On the full and global accuracy of a compact third order WENO scheme. SIAM J. Numer. Anal. **52**(5), 2335–2355 (2014)
13. A. Kurganov, D. Levy, A third-order semidiscrete central scheme for conservation laws and convection-diffusion equations. SIAM J. Sci. Comput. **22**(4), 1461–1488 (2000)
14. D. Levy, G. Puppo, G. Russo, Compact central WENO schemes for multidimensional conservation laws. SIAM J. Sci. Comput. **22**(2), 656–672 (2000)

15. X.-D. Liu, S. Osher, T. Chan, Weighted essentially non-oscillatory schemes. J. Comput. Phys. **115**(1), 200–212 (1994)
16. S. Serna, A. Marquina, Power ENO methods: a fifth-order accurate weighted power ENO method. J. Comput. Phys. **194**(2), 632–658 (2004)
17. N.K. Yamaleev, M.H. Carpenter, A systematic methodology for constructing high-order energy stable WENO schemes. J. Comput. Phys. **228**(11), 4248–4272 (2009)
18. N.K. Yamaleev, M.H. Carpenter, Third-order energy stable WENO scheme. J. Comput. Phys. **228**(8), 3025–3047 (2009)

The Application of the Boundary Element Method to the Theory of MHD Faraday Generators

Adrian Carabineanu

Abstract The problem in dimensionless variables reduces to three systems of equations for the stream function and the electric potential in three regions of a strip (the rectangular domain bounded by the electrodes and two half-strips). The singular integral equations obtained from the integral representation of the solutions and the matching conditions are disctretized and a linear system of algebraic equations is obtained. The velocity, the electric field and the generator power are calculated.

1 The MHD Faraday Generator

In the domain bounded by the electrodes, a magnetic field is transversely applied to the motion of an electrically conducting fluid flowing inside an insulated duct (Fig. 1). Electrically charged particles (ions and electrons) flowing with the fluid determine an induced electric field which drives an electric current. The electric current flowing across the electroconductive plasma between the electrodes is the *Faraday current* which is collected by the electrodes and flows in an external load circuit. It provides the main electrical output of the MHD power generator. In this paper we present a simplified version of the MHD generator theory.

2 The Boundary Value Problem

In the book [3], Chapter 7, L. Dragoş reduced the problem of MHD generators to a boundary value problem in a strip (see Fig. 2). The segments $y = 1, |x| < a$ and $y = -1, |x| < a$ represent the electrodes.

A. Carabineanu (✉)
Department of Mathematics, University of Bucharest, str. Academiei 14, Bucharest, Romania

Institute of Mathematical Statistics and Applied Mathematics of Romanian Academy, Calla 13 September 13, Bucharest, Romania
e-mail: acara@fmi.unibuc.ro

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_7

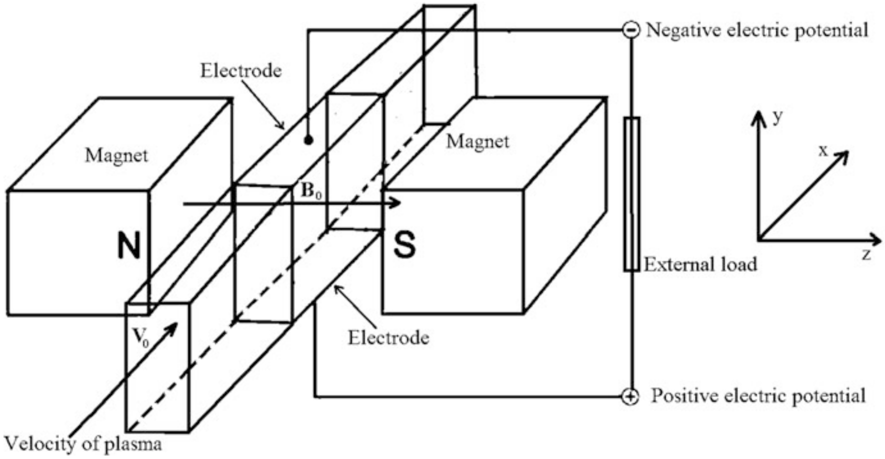


Fig. 1 MHD Faraday generator

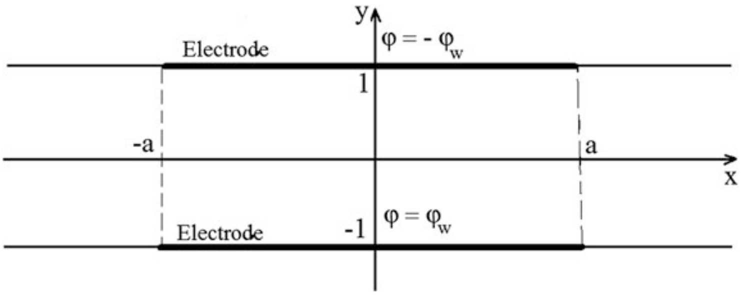


Fig. 2 Geometry of the domain in the case of plane-parallel motion

2.1 The Equations

The unknown functions are ψ, the perturbation of the stream function, φ, the electric potential and χ the harmonic conjugate of φ. We have to solve the following equations:

- in $D_{-\infty}^{-a} = \{(x, y), x < -a, -1 < y < 1\}$:

$$\Delta\chi = 0, \ \Delta\psi = 0, \tag{1}$$

– in $D_{-a}^{a} = \{(x, y), -a < x < a, -1 < y < 1\}$

$$\Delta\psi = -N\left(\frac{\partial\varphi^*}{\partial x} + \frac{\partial\psi}{\partial x}\right)_{x=-a+0}, \tag{2}$$

$$\Delta\varphi^* = N\left(\frac{\partial\varphi^*}{\partial x} + \frac{\partial\psi}{\partial x}\right)_{x=-a+0}, \ \varphi^*(x, y) = \varphi(x, y) + \varphi_w y. \tag{3}$$

– in $D_{a}^{\infty} = \{(x, y), a < x, -1 < y < 1\}$:

$$\Delta\chi = 0, \tag{4}$$

$$\Delta\psi = N\left(\frac{\partial\varphi^*}{\partial x} + \frac{\partial\psi}{\partial x}\right)_{x=a-0} - N\left(\frac{\partial\varphi^*}{\partial x} + \frac{\partial\psi}{\partial x}\right)_{x=-a+0}. \tag{5}$$

The constant N is the Stuart number.

2.2 Boundary and Matching Conditions

The boundary conditions are

$$\psi(x, \pm 1) = 0, -\infty < x < \infty, \ \varphi(x, \pm 1) = \mp\varphi_w \Longrightarrow \varphi^*(x, \pm 1) = 0, \tag{6}$$

$$\left.\begin{array}{l} \dfrac{\partial\varphi(x, \pm 1)}{\partial y} = 0, |x| > a \\ \lim\limits_{x\to\pm\infty} grad\varphi = 0 \end{array}\right\} \Longrightarrow \chi(x, \pm 1) = 0, |x| > a. \tag{7}$$

The matching conditions are

$$\frac{\partial\chi(-a, y)}{\partial y} - \frac{\partial\varphi^*(-a, y)}{\partial x} - \frac{\partial\psi(-a, y)}{\partial x} = 0, \tag{8}$$

$$\frac{\partial\chi(a, y)}{\partial y} - \frac{\partial\varphi^*(a, y)}{\partial x} - \frac{\partial\psi(a, y)}{\partial x} = 0, \tag{9}$$

$$[\varphi]_{x=\pm a} = 0 \Longrightarrow \varphi^*(\pm a, y) + \int_0^y \frac{\partial\chi(\pm a, y)}{\partial x} dy = \varphi_w y. \tag{10}$$

3 Singular Equations

In the integral representations we pass to limit for $x \to \pm a \pm 0$, we use Plemelj formulas and we take into account that $\dfrac{\partial g(\xi, y; \xi, \eta)}{\partial \xi} = 0$.

For $x \to -a - 0$ we have

$$\frac{1}{2}\chi(-a, y) = \int_{-1}^{1} \frac{\partial \chi(-a, \eta)}{\partial \xi} g(-a, y; -a, \eta) d\eta, \tag{11}$$

$$\frac{1}{2}\psi(-a, y) = \int_{-1}^{1} \frac{\partial \psi(-a, \eta)}{\partial \xi} g(-a, y; -a, \eta) d\eta. \tag{12}$$

For $x \to -a + 0$ we have

$$\frac{1}{2}\psi\ (-a, y) = \int_{-1}^{1} \left[\frac{\partial \psi(a, \eta)}{\partial \xi} g(-a, y; a, \eta) - \frac{\partial g(-a, y; a, \eta)}{\partial \xi} \psi(a, \eta)\right] d\eta - \dots$$

$$- \int_{-1}^{1} \frac{\partial \psi(-a, \eta)}{\partial \xi} g(-a, y; -a, \eta) d\eta + \dots$$

$$N \int_{-1}^{1} \left(\frac{\partial \varphi^*(-a, \eta)}{\partial \xi} + \frac{\partial \psi(-a, \eta)}{\partial \xi}\right) G_{-a}^{a}(-a, y; \eta) d\eta, \tag{13}$$

$$\frac{1}{2}\varphi^*\ (-a, y) = \int_{-1}^{1} \left[\frac{\partial \varphi^*(a, \eta)}{\partial \xi} g(-a, y; a, \eta) - \frac{\partial g(-a, y; a, \eta)}{\partial \xi} \varphi^*(a, \eta)\right] d\eta - \dots$$

$$- \int_{-1}^{1} \frac{\partial \varphi^*(-a, \eta)}{\partial \xi} g(-a, y; -a, \eta) \dots$$

$$+ N \int_{-1}^{1} \left(\frac{\partial \varphi^*(-a, \eta)}{\partial \xi} + \frac{\partial \psi(-a, \eta)}{\partial \xi}\right) G_{-a}^{a}(-a, y; \eta) d\eta, \tag{14}$$

$$\frac{1}{2}\varphi^*\ (-a, y) = \int_{-1}^{1} \left[\frac{\partial \varphi^*(a, \eta)}{\partial \xi} g(-a, y; a, \eta) - \frac{\partial g(-a, y; a, \eta)}{\partial \xi} \varphi^*(a, \eta)\right] d\eta - \dots$$

$$- \int_{-1}^{1} \frac{\partial \varphi^*(-a, \eta)}{\partial \xi} g(-a, y; -a, \eta) \dots$$

$$- N \int_{-1}^{1} \left(\frac{\partial \varphi^*(-a, \eta)}{\partial \xi} + \frac{\partial \psi(-a, \eta)}{\partial \xi}\right) G_{-a}^{a}(-a, y; \eta) d\eta. \tag{15}$$

For $x \to a - 0$ we have

$$\frac{1}{2}\psi\,(a,y) = \int_{-1}^{1} \frac{\partial\psi(a,\eta)}{\partial\xi} g(a,y;a,\eta)d\eta - \dots$$

$$-\int_{-1}^{1}\left[\frac{\partial\psi(-a,\eta)}{\partial\xi} g(a,y;-a,\eta) - \frac{\partial g(a,y;-a,\eta)}{\partial\xi}\psi(-a,\eta)\right]d\eta + \dots$$

$$N\int_{-1}^{1}\left(\frac{\partial\varphi^*(-a,\eta)}{\partial\xi} + \frac{\partial\psi(-a,\eta)}{\partial\xi}\right)G_{-a}^{a}(a,y;\eta)d\eta, \tag{16}$$

$$\frac{1}{2}\varphi^*\,(a,y) = \int_{-1}^{1} \frac{\partial\varphi^*(a,\eta)}{\partial\xi} g(a,y;a,\eta)d\eta - \dots$$

$$-\int_{-1}^{1}\left[\frac{\partial\varphi^*(-a,\eta)}{\partial\xi} g(a,y;-a,\eta) - \frac{\partial g(a,y;-a,\eta)}{\partial\xi}\varphi^*(-a,\eta)\right]d\eta - \dots$$

$$-N\int_{-1}^{1}\left(\frac{\partial\varphi^*(-a,\eta)}{\partial\xi} + \frac{\partial\psi(-a,\eta)}{\partial\xi}\right)G_{-a}^{a}(a,y;\eta)d\eta. \tag{17}$$

For $x \to a + 0$ we have

$$\frac{1}{2}\chi\,(a,y) = -\int_{-1}^{1} \frac{\partial\chi(a,\eta)}{\partial\xi} g(x,y;-a,\eta)d\eta,\ a < x, \tag{18}$$

$$\frac{1}{2}\psi\,(a,y) = -\int_{-1}^{1} \frac{\partial\psi(a,\eta)}{\partial\xi} g(a,y;a,\eta)d\eta - \dots$$

$$-N\int_{-1}^{1}\left(\frac{\partial\varphi^*(-a,\eta)}{\partial\xi} + \frac{\partial\psi(-a,\eta)}{\partial\xi} - \dots\right.$$

$$\left.-\frac{\partial\varphi^*(a,\eta)}{\partial\xi} - \frac{\partial\psi(a,\eta)}{\partial\xi}\right)G_{a}^{\infty}(a,y;\eta)d\eta. \tag{19}$$

We denoted the Green function for the strip

$$g\,(x,y;\xi,\eta) = -\frac{1}{4\pi}\ln\frac{\cosh\frac{\pi}{2}(x-\xi) + \cos\frac{\pi}{2}(y+\eta)}{\cosh\frac{\pi}{2}(x-\xi) - \cos\frac{\pi}{2}(y-\eta)}, \tag{20}$$

and $G_{-a}^{a}(x,y;\eta) = \int_{-a}^{a} g(x,y;\xi,\eta)d\xi$, $G_{a}^{\infty}(x,y;\eta) = \int_{a}^{\infty} g(x,y;\xi,\eta)d\xi$.

3.1 The Unknown Functions

We have 8 singular integral equations (with logarithmic singularity) and 4 matching conditions for determining the 12 unknown functions $\chi(-a,y)$, $\dfrac{\partial\chi(-a,y)}{\partial x}$, $\chi(a,y)$, $\dfrac{\partial\chi(a,y)}{\partial x}$, $\varphi^*(-a,y)$, $\dfrac{\partial\varphi^*(-a,y)}{\partial x}$, $\varphi^*(a,y)$, $\dfrac{\partial\varphi^*(a,y)}{\partial x}$
and $\psi(-a,y)$, $\dfrac{\partial\psi(-a,y)}{\partial x}$, $\psi(a,y)$, $\dfrac{\partial\psi(a,y)}{\partial x}$.

Some theoretical results enable us to presume the following behaviour:

$$\frac{\partial\chi(-a,y)}{\partial x}=\frac{\chi_x^*(-a,y)}{\sqrt{1-y^2}},\ \frac{\partial\chi(a,y)}{\partial x}=\frac{\chi_x^*(a,y)}{\sqrt{1-y^2}},\ \frac{\partial\varphi^*(-a,y)}{\partial x}=\frac{\varphi_x^*(-a,y)}{\sqrt{1-y^2}},$$
$$\frac{\partial\varphi^*(a,y)}{\partial x}=\frac{\varphi_x^*(a,y)}{\sqrt{1-y^2}},\ \frac{\partial\psi(-a,y)}{\partial x}=\frac{\psi_x^*(-a,y)}{\sqrt{1-y^2}},\ \frac{\partial\psi(a,y)}{\partial x}=\frac{\psi_x^*(a,y)}{\sqrt{1-y^2}},$$

where $\chi_x^*(-a,y)$, $\chi_x^*(a,y),\varphi_x^*(-a,y),\varphi_x^*(a,y),\psi_x^*(-a,y),\psi_x^*(a,y)$ are bounded for $-1\le y\le 1$.

4 Discretization

We split the segments $x=-a,-1\le y\le 1$ and $x=a,-1\le y\le 1$ into n panels $\left\{\Gamma_j^{(-a)}\right\}_{j=1,\dots,n}$ respectively $\left\{\Gamma_j^{(a)}\right\}_{j=1,\dots,n}$ such that

$$\Gamma_j^{(-a)}=\left\{(-a,y)\,;-1+\eta_j^{(1)}\le y\le -1+\eta_{j+1}^{(1)}\right\},\eta_j^{(1)}=-1+\frac{2(j-1)}{n},$$
$$\Gamma_j^{(a)}=\left\{(a,y)\,;-1+\frac{2(j-1)}{n}\le y\le -1+\frac{2j}{n}\right\}.$$

In every panel we take the midpoints

$$\left(-a,\eta_j\right)=-1+\frac{2j-1}{n},j=1,\dots,n$$

We approximate the functions with piecewise constant functions

$$\int_{-1}^{1}\frac{\partial\psi(a,\eta)}{\partial\xi}g(-a,\eta_i;a,\eta)d\eta\approx$$
$$\approx\sum_{j=1}^{n}\psi_\xi^*(-a,\eta_j)g(-a,\eta_i;-a,\eta_j)\left(\arcsin\eta_{j+1}^{(1)}-\arcsin\eta_j^{(1)}\right),$$

$$\int_{-1}^{1} \frac{\partial g(-a,y;a,\eta)}{\partial \xi}\psi(a,\eta)d\eta \approx \frac{1}{n}\sum_{j=1}^{n}\psi(-a,\eta_j)\frac{\partial g(-a,\eta_i;a,\eta_j)}{\partial \xi},$$

$$\int_{\Gamma_j^{(-a)}} \frac{\partial \chi(-a,\eta)}{\partial \xi} g(-a,\eta_i;-a,\eta)d\eta = \int_{\Gamma_j^{(-a)}} \frac{\chi_\xi^*(-a,\eta)}{\sqrt{1-\eta^2}} g(-a,\eta_i;-a,\eta)d\eta \approx$$

$$\approx \chi_\xi^*(-a,\eta_j)\int_{\Gamma_j^{(-a)}} \frac{g(-a,\eta_i;-a,\eta)}{\sqrt{1-\eta^2}}d\eta \approx$$

$$\approx \chi_\xi^*(-a,\eta_j)g(-a,\eta_i;-a,\eta_j)\left(\arcsin\eta_{j+1}^{(1)} - \arcsin\eta_j^{(1)}\right), i \neq j,$$

$$\int_{\Gamma_i^{(-a)}} \frac{\chi_\xi^*(-a,\eta)}{\sqrt{1-\eta^2}} g(-a,\eta_i;-a,\eta)d\eta \approx \chi_\xi^*(-a,\eta_i)\int_{\Gamma_i^{(-a)}} \frac{g(-a,\eta_i;-a,\eta)}{\sqrt{1-\eta^2}}d\eta \approx$$

$$\approx \chi_\xi^*(-a,\eta_i)\left(\int_{-1}^{1} \frac{g(-a,\eta_i;-a,\eta)}{\sqrt{1-\eta^2}}d\eta - \sum_{j\neq i,j=1}^{n}\int_{\Gamma_j^{(-a)}} \frac{g(-a,\eta_i;-a,\eta)}{\sqrt{1-\eta^2}}d\eta\right) \approx$$

$$\chi_\xi^*(-a,\eta_i)\left(\frac{\pi}{n}\sum_{\alpha=1}^{n} g\left(-a,\eta_i;-a,\cos\frac{(2\alpha-1)\pi}{2n}\right) - \right.$$

$$\left. - \sum_{j\neq i,j=1}^{n} g(-a,\eta_i;-a,\eta_j)\left(\arcsin\eta_{j+1}^{(1)} - \arcsin\eta_j^{(1)}\right)\right).$$

$$\int_{-1}^{1} \frac{\partial \psi(-a,\eta)}{\partial \xi} G_{-a}^{a}(a,y;\eta)d\eta \approx$$

$$\approx \sum_{j=1}^{n}\psi_\xi^*(-a,\eta_j)\left(\arcsin\eta_{j+1}^{(1)} - \arcsin\eta_j^{(1)}\right)\frac{G_{-a}^{a}(a,y;\eta_j^{(1)}) + G_{-a}^{a}(a,y;\eta_{j+1}^{(1)})}{2}.$$

From the matching condition

$$\varphi^*(\pm a,y) + \int_0^y \frac{\partial \chi(\pm a,y)}{\partial x}dy = \varphi_w y$$

we get the linear equations

$$\varphi^*(\pm a,\eta_j) - \varphi^*(\pm a,\eta_{j-1}) +$$

$$+\frac{\chi_\xi^*(\pm a,\eta_j) + \chi_\xi^*(\pm a,\eta_{j-1i})}{2}\left(\arcsin\eta_j^{(1)} - \arcsin\eta_{j-1}^{(1)}\right) = \frac{2\varphi_w}{n},\ j = 2,..n,$$

$$\varphi^*(\pm a,\eta_1^{(1)}) + \chi_\xi^*(\pm a,\eta_1)\arcsin\eta_1^{(1)} = \frac{\varphi_w}{n}$$

and so on.

4.1 The Linear System

After discretizing the integral equations and matching equations we obtain a linear non-singular algebraic system. The unknowns are

$\chi(-a,\eta_i),\ \chi^*_\xi(-a,\eta_i),\ \chi(a,\eta_i),\ \chi^*_\xi(a,\eta_i),\ \varphi^*(-a,\eta_i),\ \varphi^*_\xi(-a,\eta_i),\ \varphi^*(a,\eta_i),$ $\varphi^*_\xi(a,\eta_i),\ \psi(-a,\eta_i),\ \psi^*_\xi(-a,\eta_i),\ \psi(a,\eta_i),\ \psi^*_\xi(a,\eta_i), i = 1,\dots,n.$

5 The Velocity and the Electric Current

We denote by (x_l, y_k) the coordinates of the points of a grid in the strip $-1 < y < 1, x \in \mathbb{R}$. The velocity in the grid points is:

$$V_x(x_l,y_k) = 1 + \frac{\partial\psi(x_l,y_k)}{\partial y},\ V_y(x_l,y_k) = -\frac{\partial\psi(x_l,y_k)}{\partial x}.$$

The electric current for $|x| > a$ is

$$\frac{J_x(x_l,y_k)}{Rm} = -\frac{\partial\chi(x_l,y_k)}{\partial y},\ \frac{J_y(x_l,y_k)}{Rm} = \frac{\partial\chi(x_l,y_k)}{\partial x}.$$

and for $|x| < a$

$$\frac{J_x(x_l,y_k)}{Rm} = -\frac{\partial\chi(x_l,y_k)}{\partial y} - \frac{\partial\psi(x_l,y_k)}{\partial y},\ \frac{J_y(x_l,y_k)}{Rm} = \frac{\partial\chi(x_l,y_k)}{\partial x} + \frac{\partial\psi(x_l,y_k)}{\partial y} - 1$$

where Rm is the magnetic Reynolds number. In Fig. 3 we present the nondimensional velocity, streamlines and the electric current for some values of the length of electrodes, electric potential on electrodes and Stuart number.

6 The Output Power

The useful power developed by the MHD generator is

$$W = -\varphi_w Rm \int_{-a}^{a} \left[J_y(x,-1) + J_y(x,1)\right] dx.$$

In Fig. 4 we present the useful output power W/φ_w for various values of the length of electrodes a and N.

We notice that for $N = 0$ the numerical results almost coincide with the analytical results obtained using the results from papers [1, 2].

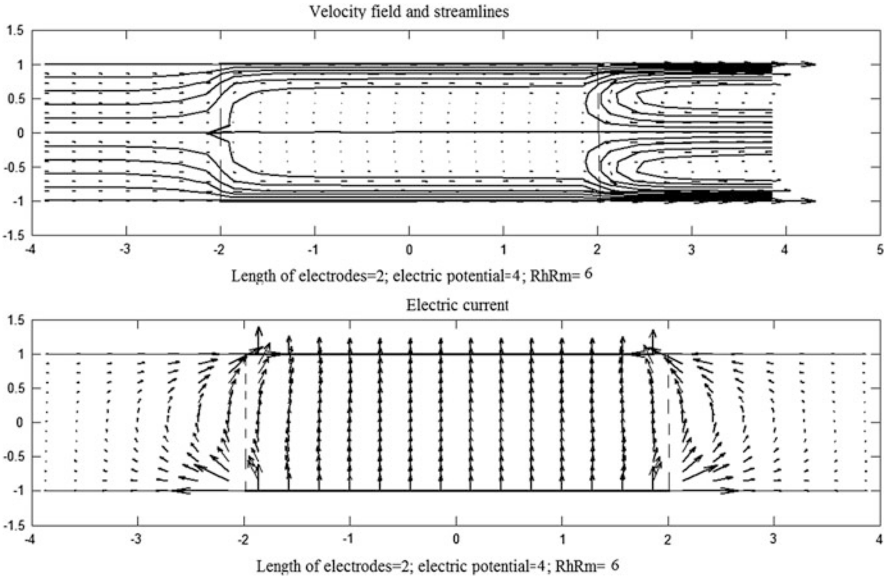


Fig. 3 Velocity, streamlines and electric current

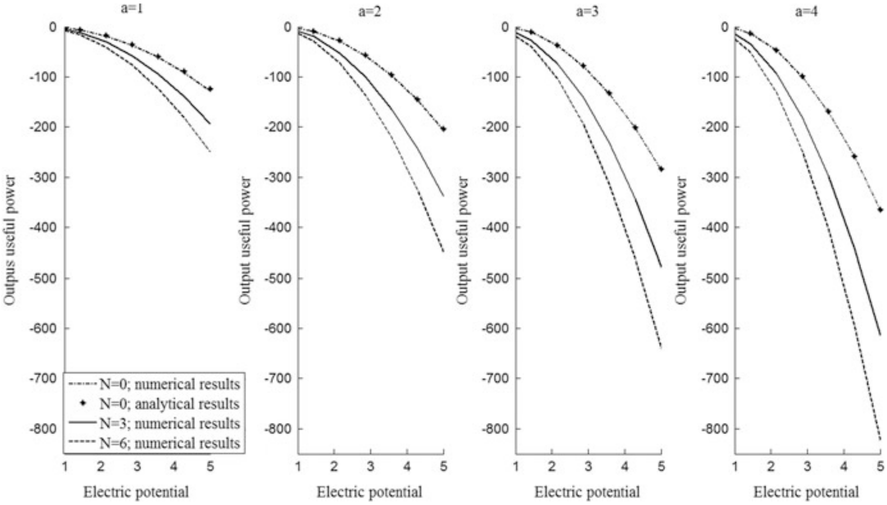


Fig. 4 Useful output power

References

1. A. Carabineanu, Numerical calculation of the output power of a mhd generator. INCAS BULLETIN **6**(4), 15–22 (2014)
2. A. Carabineanu, A simplified mathematical theory of MHD power generators. An. St. Univ. Ovidius Constanta **23**(3), 29–39 (2015)
3. L. Dragoş, *Magneto-Fluid Dynamics* (Editura Academiei/Abacus Press, Bucureşti/Tunbridge Wells, 1975)

Part II
Finite Element Methods

How to Avoid Mass Matrix for Linear Hyperbolic Problems

Rémi Abgrall, Paola Bacigaluppi, and Svetlana Tokareva

Abstract We are interested in the numerical solution of linear hyperbolic problems using continuous finite elements of arbitrary order. It is well known that this kind of methods, once the weak formulation has been written, leads to a system of ordinary differential equations in $\mathbb{R}^N$, where N is the number of degrees of freedom. The solution of the resulting ODE system involves the inversion of a sparse mass matrix that is not block diagonal. Here we show how to avoid this step, and what are the consequences of the choice of the finite element space. Numerical examples show the correctness of our approach.

1 Introduction

We are interested in the numerical approximation of the hyperbolic problem

$$\frac{\partial u}{\partial t} + \text{ div } \mathbf{f}(\mathbf{x}, u) = 0 \qquad \mathbf{x} \in \Omega \subset \mathbb{R}^d \tag{1a}$$

by means of a finite element like technique. In this paper, we focus on the linear case where $\mathbf{f}(\mathbf{x}, u) = \mathbf{a}(\mathbf{x})u$. The vector field $\mathbf{a}$ may depend on the spatial location $\mathbf{x}$. The problem (1a) is also supplemented with initial and boundary conditions:

$$u(\mathbf{x}, 0) = u_0(\mathbf{x}) \tag{1b}$$

and

$$u(\mathbf{x}, t) = g(\mathbf{x}) \text{ if } \mathbf{x} \in \partial\Omega, t \geq 0. \tag{1c}$$

R. Abgrall (✉) • P. Bacigaluppi • S. Tokareva
Institute of Mathematics, University of Zürich, Winterthurerstrasse 190, CH 8057 Zürich, Switzerland
e-mail: remi.abgrall@math.uzh.ch; paola.bacigaluppi@math.uzh.ch; svetlana.tokareva@math.uzh.ch

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_8

Obviously, (1c) has to be understood in the weak sense, i.e. that $u = g$ on the inflow characteristics.

The physical space is covered by a conformal tessellation $\mathscr{T}$. For ease of exposition, we assume that

$$\Omega = \cup_{K \in \mathscr{T}} K.$$

The solution of the problem is approximated by an element of the space V^h defined by:

$$V^h = \{\mathbf{u}^h \in C^0(\Omega) \text{ such that for any } K, \mathbf{u}^h_{|K} \text{ is a polynomial of degree } r\}.$$

We denote by $\mathbb{P}^r$ the set of polynomials of degree r. In this paper, we consider $r = 1, 2$ only.

It is well known that any finite element technique applied to (1a) will lead to a formulation of the type

$$M\frac{dU}{dt} + F = 0$$

where U denotes the vector of degrees of freedom, F is an approximation of the term div $\mathbf{f}$ and M is a mass matrix. In the case of continuous elements, this matrix is sparse but not block diagonal, contrarily to what happens for the Discontinuous Galerkin methods where the global continuity requirement is not made. Hence, in order to use any standard ODE solver, we need to invert M. This is considered cumbersome by many practitioners and this has been, in our opinion, one of the factors that has led to supremacy of DG methods in the current development of high order schemes.

Several researchers have proposed methods that avoid this step. More precisely, their methods are designed in such a way that the actual mass matrix is diagonal, so that the problem amounts to finding a "good" lumping integration formula. The first work we are aware of in that direction is [4], where the wave equation is considered, and the finite element space is made of functions belonging to a subspace of $\mathbb{P}^{k+1}$ that contains $\mathbb{P}^k$. This amounts to adding one degree of freedom to the "natural" quadratic elements. This work has been followed, in the same spirit, by [9] where higher accuracy could be obtained. However, the elements become more and more complex and, what is even more important, the stability condition on the time step becomes dramatically restrictive.

In these notes we describe some preliminary results about a new method for which no inversion of the mass matrix is needed, while a typical finite element approximation can be kept for the description of the divergence term. In this approach, there is no need to change the degrees of freedom. The method presented here can be seen as an extension of [10] where only $\mathbb{P}^1$ elements and second order approximation in time have been considered.

The rest of the paper is organized as follows. In the first section, we describe the approximation of the divergence term of (1a). These are classical stabilized finite element methods. In the second section, we describe and somewhat justify our approach. The last section provides numerical examples that justify the correctness of our approach. A more involved analysis and description will be made elsewhere. We conclude by giving some perspectives.

2 Description of the Scheme

We start by describing the two spatial approximations we consider, then explain how to avoid the mass matrix inversion. We are given a triangulation of $\mathbb{R}^d$. Here we assume $d = 2$, but the discussion is general. The elements are denoted by K and assumed to be simplices. In each element, we assume that the solution is approximated by a polynomial of degree r and that the approximation is globally continuous. Let us denote the approximate solution by u^h. The function u^h is fully defined by its control parameter u_σ at all the degrees of freedom σ. We define by $\mathscr{S}$ the set of degrees of freedom, so that

$$u^h = \sum_{\sigma \in \mathscr{S}} u_\sigma \varphi_\sigma.$$

We denote by $V_h = \text{span}\,\{\varphi_\sigma, \sigma \in \mathscr{S}\}$. For now, we can think of u_σ as the value of u^h at σ and thus φ_σ is the Lagrange basis, but we will need slightly less conventional approximation later.

We assume that we have a good integrator of the steady version of (1), and that this scheme writes: for any degree of freedom σ, u^h satisfies:

$$\sum_{K \ni \sigma} \Phi_\sigma^{K,\mathbf{x}}(u^h) = 0.$$

Examples are given by:

1. The SUPG residual, [7, 8]:

$$\begin{aligned} \Phi_\sigma^{\mathbf{x}}(u^h) = &\int_{\partial K} \varphi_\sigma \mathbf{f}(u^h) \cdot \mathbf{n} \, d\ell - \int_K \nabla \varphi_\sigma \cdot \mathbf{f}(u^h) \, d\mathbf{x} \\ &+ h_K \int_K \Big(\nabla_u \mathbf{f}(u^h) \cdot \nabla \varphi_\sigma \Big) \tau \Big(\nabla_u \mathbf{f}(u^h) \cdot \nabla u^h \Big) \, d\mathbf{x} \end{aligned} \tag{2}$$

with $\tau > 0$. We take:

$$(h_K \tau)^{-1} = \sum_{\sigma \in K} |\overline{\mathbf{a}}_K \cdot \nabla \varphi_\sigma|$$

where $\overline{\mathbf{a}}_K$ is the value of $\mathbf{a}$ at the centroid of K.

2. The Galerkin scheme with jump stabilization [3]:

$$\Phi_\sigma^{\mathbf{x}}(u^h) = \int_{\partial K} \varphi_\sigma \mathbf{f}(u^h) \cdot \mathbf{n} \, d\ell - \int_K \nabla\varphi_\sigma \cdot \mathbf{f}(u^h) \, d\mathbf{x} \\ + \sum_{edges} \Gamma h_e^2 \int_e [\nabla u] \cdot [\nabla \varphi_\sigma]^+ \, d\ell \tag{3}$$

   with $\Gamma > 0$. Here, since the mesh is conformal, any edge (or face in 3D) is the intersection of the element K and an other element denoted by K^+. We define $[\nabla u] = \nabla u_{|K} - \nabla u_{|K^+}$ and $[\nabla\varphi_\sigma]^+ = (\varphi_\sigma)_{|K}$. Here, we have taken $\Gamma = \max(\overline{\mathbf{a}}_K, \overline{\mathbf{a}}_{K^+})$. See [3] for more details.

This streamline formulation implies formally that the exact solution cancels the residuals. In the case of the stabilisation by jumps, we can only write that

$$\Phi_\sigma^K = \int_K \varphi_\sigma \text{div } \mathbf{f}(u) d\mathbf{x} + R_\sigma(u^h)$$

where $\sum_{\sigma\in K} R_\sigma(u^h) = 0$. The additional term R_σ is non-zero, except for the exact solution unless this solution has continuous normal gradients. For steady solutions, both methods can be shown to converge as $h^{k+1/2}$, see [3, 8] for more details.

2.1 Formulation for Unsteady Problems

We use a deferred correction (DeC) approach. We start from the ODE:

$$\frac{dy}{dt} = f(y,t), \qquad y(0) = y_0. \tag{4}$$

We follow the main ideas of [6]. Between t_n and t_{n+1}, the solution of (4) satisfies

$$y(t) = y(t_n) + \int_{t_n}^t f(y(s), s) ds.$$

Given $0 = \xi_0 < \xi_1 < \ldots < \xi_l < \ldots < \xi_{M+1} = 1$, and we consider the times $t_{n,l} = t_n + \xi_l \Delta t$ with $\Delta t = t_{n+1} - t_n$. If we know $f_{n,l} \approx f(y(t_{n,l}), t_{n,l})$, we can consider the Lagrange interpolant $\mathscr{I}_{M+1}$ of f with data given by $(t_{n,l}, f_{n,l})$, therefore we get the approximation:

$$y_{n,l} = y_{n,0} + \int_{t_n}^{t_{n,l}} \mathscr{I}_{M+1}[f(y(\,.\,),\,.\,)](t_n + \xi\Delta t) d\xi.$$

This is in general a non-linear implicit equation.

The idea of the DeC method is to consider the first order scheme, for $M \geq l \geq 1$:

$$y_{n,l} = y_{n,l-1} + \alpha_l \Delta t f(y(t_{n,l-1}), t_{n,l-1}), \qquad y_{n,0} \approx y(t_n)$$

where $\alpha_l = \xi_l - \xi_{l-1}$. Then, we introduce the vector $v = (y_{n,1}, \dots, y_{n,M+1})^T$. The first order scheme can be rewritten as $L^1(v) = 0$ where

$$L^1(v) = \begin{pmatrix} y_{n,1} - y_{n,0} - \Delta t \int_0^{\xi_1} \mathscr{I}_0[f(y(\,.\,),\,.\,)](t_n + \xi \Delta t) d\xi \\ \vdots \\ y_{n,l} - y_{n,0} - \Delta t \int_0^{\xi_l} \mathscr{I}_0[f(y(\,.\,),\,.\,)](t_n + \xi \Delta t) d\xi \\ \vdots \\ y_{n,M+1} - y_{n,0} - \Delta t \int_0^{\xi_M} \mathscr{I}_0[f(y(\,.\,),\,.\,)](t_n + \xi \Delta t) d\xi \end{pmatrix}$$

where $\mathscr{I}_0$ is the first order interpolant of f: for $1 \leq l \leq M+1$,

$$\mathscr{I}_0[f(y(\,.\,),\,.\,)](s) = f(y_{n,l-1}, t_{n,l-1}) \qquad \text{for } s \in [t_{n,l-1}, t_{n,l}[.$$

Note that $L^1(v) = 0$ can be solved *explicitely*.

Similarly, we define L^2 by:

$$L^2(v) = \begin{pmatrix} y_{n,1} - y_{n,0} - \Delta t \int_0^{\xi_1} \mathscr{I}_{M+1}[f(y(\,.\,),\,.\,)](t_n + \xi \Delta t) d\xi \\ \vdots \\ y_{n,l} - y_{n,0} - \Delta t \int_0^{\xi_l} \mathscr{I}_{M+1}[f(y(\,.\,),\,.\,)](t_n + \xi \Delta t) d\xi \\ \vdots \\ y_{n,M+1} - y_{n,0} - \Delta t \int_0^{\xi_M} \mathscr{I}_{M+1}[f(y(\,.\,),\,.\,)](t_n + \xi \Delta t) d\xi \end{pmatrix}.$$

The DeC formulation is defined as follows:

1. $v^0 = (y_n, \dots y_n)^T$ and $y_{n,0} = y_n$,
2. For $k = 1, \dots M+1$, v^k is defined as

$$L^1(v^k) = L^1(v^{k-1}) - L^2(v^{k-1})$$

Since L^1 is explicit, the method is completely explicit. One can show that $L^2 - L^1 = O(\Delta t)$ so that the scheme is $(M+1)$-th order accurate.

Similar to what is done for ODEs, we could integrate (1) in time and get:

$$u(\mathbf{x}, t_{n+1}) = u(\mathbf{x}, t_n) + \int_{t_n}^{t_{n+1}} \text{div } \mathbf{f}(u(x,s)) ds,$$

This can be approximated by

$$\begin{aligned} u(\mathbf{x}, t_n + \xi_i \Delta t) &\approx u(\mathbf{x}, t_n) + \int_0^{\xi_i} \text{div } \mathscr{I}_{r+1}[\mathbf{f}(u(\mathbf{x}, \ .\ )](t_n + \xi \Delta t) ds \\ &= \Delta t \sum_{l=0}^{r} \omega_l^i \text{div } \mathbf{f}(u(\mathbf{x}, \xi_j) ds \end{aligned} \tag{5}$$

$\mathscr{I}_{r+1}[\mathbf{f}(u(\mathbf{x}, \ .\ ))]$ is the Lagrange interpolant of $\mathbf{f}(u(\mathbf{x}, \ .\ ))$ at the points $\{t_n, , \dots, \xi_i \Delta t, \dots, t_{n+1}\}$ and ω_l^i are the weights.

This suggests the algorithm we describe now. For any $V \in V_h^M$, $V^\sigma = (V_1^\sigma, \dots, V_{M+1}^\sigma)^T$ is a vector of control parameters at the degree of freedom $\sigma \in \mathscr{S}$: $V = \sum_{\sigma \in \mathscr{S}} V^\sigma \varphi_\sigma$. Then, we can consider the following deferred correction approximation: we introduce $t_{n,i} = t_n + \xi_i(t_{n+1} - t_n)$ so that $t_{n,0} = t_n$ and $t_{n,r+1} = t_{n+1}$, and define

1. for any $\sigma \in \mathscr{S}$, the operator L_σ^1 as

$$L_\sigma^1(V_1, \dots, V_{r+1}) = \begin{pmatrix} |C_\sigma|(V_{r+1}^\sigma - V_0^\sigma) + \sum\limits_{K \ni \sigma} \int_{t_{n,0}}^{t_{n,r+1}} \mathscr{I}_0[\Phi_\sigma^{\mathbf{x}}]\big(t_n + s\Delta t\big)\, ds \\ |C_\sigma|(V_r^\sigma - V_0^\sigma) + \sum\limits_{K \ni \sigma} \int_{t_{n,0}}^{t_{n,r}} \mathscr{I}_0[\Phi_\sigma^{\mathbf{x}}]\big(t_n + s\Delta t)\big)\, ds \\ \vdots \\ |C_\sigma|(V_1^\sigma - V_0^\sigma) + \sum\limits_{K \ni \sigma} \int_{t_{n,0}}^{t_{n,1}} \mathscr{I}_0[\Phi_\sigma^{\mathbf{x}}]\big(t_n + s\Delta t\big)\, ds \end{pmatrix} \tag{6a}$$

Here, $V_0^\sigma = (u_\sigma^n, \dots, u^n|\sigma)^T \in \mathbb{R}^M$.

2. and the operator L_σ^2 as

$$L_\sigma^2(V_1, \dots, V_{r+1}) = \begin{pmatrix} \sum\limits_{K \ni \sigma} \left(\int_K \Psi_\sigma\big(V_{r+1} - V_0\big)\, dx + \int_{t_{n,0}}^{t_{n,r+1}} \mathscr{I}_{r+1}[\Phi_\sigma^{\mathbf{x}}]\big(t_n + s\Delta t\big)\, ds \right. \\ \sum\limits_{K \ni \sigma} \left(\int_K \Psi_\sigma\big(V_r - V_0\big)\, dx + \int_{t_{n,0}}^{t_{n,r}} \mathscr{I}_{r+1}[\Phi_\sigma^{\mathbf{x}}]\big(t_n + s\Delta t\big)\, ds \right. \\ \vdots \\ \sum\limits_{K \ni \sigma} \left(\int_K \Psi_\sigma\big(V_1 - V_0\big)\, dx + \int_{t_{n,0}}^{t_{n,1}} \mathscr{I}_{r+1}[\Phi_\sigma^{\mathbf{x}}]\big(t_n + s\Delta t\big)\, ds \right. \end{pmatrix} \tag{6b}$$

Last, we define the operators L^1 and L^2 on the finite element set V_h as

$$L^1 = (L^1_\sigma)_{\sigma\in\mathscr{S}}, \qquad L^2 = (L^1_\sigma)_{\sigma\in\mathscr{S}}.$$

The step of the method between t_n and t_{n+1} is defined as follows.

1. Knowing u^n_σ, we set $V^0_\sigma = (u^n_\sigma, \dots, u^n_\sigma)$.
2. For $k = 1, \dots, M$, we construct V^k as the solution of

$$L^1(V^{k+1}) = L^1(V^k) - L^2(V^k).$$

3. Then we define u^{n+1}_σ as

$$u^{n+1}_\sigma = (V^{r+1}_\sigma)^M.$$

This method provides a decent approximation of the solution because one can show [1] that, for the L^2 norm,

$$||L^1 - L^2|| \leq C\Delta t, \tag{7}$$

where the constant C depends only on the mesh. Then, using standard results for deferred correction methods, one can show that we have an $(r+1)$-th order accurate scheme if $M = r + 1$, provided L^1 is invertible. The overall cost is not larger than a standard Runge-Kutta method.

Let us now have a look at the invertibility of L^1. Not every finite element approximation can work. The reason is that we have not yet specified what should be the parameters C_σ in relation (6a). It is easy to see that we must have

$$C_\sigma = \int_\Omega \varphi_\sigma d\mathbf{x},$$

and in order that L^1 be invertible, we need $C_\sigma > 0$. For $\mathbb{P}^1$ elements, there is no problem because the basis functions are positive, but it is well known that this condition is not met for higher order finite elements. For example, in the case of two-dimensional quadratic Lagrange interpolation, we have six basis functions. Three of them are associated to the vertices, and it is well known that their integral vanishes, so that in the end $C_\sigma = 0$ for the vertices. For other finite elements, we can have $C_\sigma < 0$.

In order to circumvent this restriction, and since we are interested in the approximation order and *not on the practical representation, i.e. the physical meaning of the degrees of freedom*, a simple way is to replace classical Lagrange elements of degree r by their Bezier counterparts. If

$$\left(\sum_{j=1}^{d+1} x_j\right)^r = \sum_{\sum_{k=1}^{d+1} j_k = r} \theta^r_{j_1..j_{d+1}} {x_1}^{j_1} \dots x_{j_{d+1}}$$

is the binomial expansion, then the Bezier polynomials are simply

$$B^r_{j_1 \cdots j_{d+1}} = \theta^r_{j_1 \cdots j_{d+1}} \Lambda_1^{j_1} \dots \Lambda_{j_{d+1}}$$

where the Λ_j are the standard barycentric coordinates. Since

$$\int_K B^r_{j_1 \cdots j_{d+1}}(\mathbf{x}) d\mathbf{x} > 0,$$

and since this family is a basis of $\mathbb{P}^r$, there are no more problems. In the simulations done in this paper, we have chosen $\mathbb{P}^1$ elements (i.e. Bézier of degree 1), and quadratic Bézier elements. Note that this kind of approximation has already been used for steady problems [2], and has some links with isogeometrical analysis [5], but for completely different reasons.

3 Numerical Illustrations

3.1 Parameters

In the numerical experiments we present, we have chosen a temporal scheme that is third order in time. It is based on the Lagrange interpolation in [0, 1], where the data are given at the points $t = 0, \frac{1}{2}$ and 1. This results in the following formula that defines the operator L^2:

$$\int_0^{1/2} \mathscr{I}_2(f) ds = \frac{5}{24} f(0) + \frac{1}{3} f(\frac{1}{2}) - \frac{1}{24} f(1)$$
$$\int_0^{1} \mathscr{I}_2(f) ds = \frac{1}{6} f(0) + \frac{4}{6} f(\frac{1}{2}) + \frac{1}{6} f(1)$$

We have used the same temporal scheme for $\mathbb{P}^1$ and $\mathbb{B}^2$ elements.

3.2 Simulations

The velocity field at (x, y) is given by $\mathbf{a} = 2\pi(-y, x)$. The initial condition is given by:

$$u_0(x, y) = e^{-40(x^2+y^2)}.$$

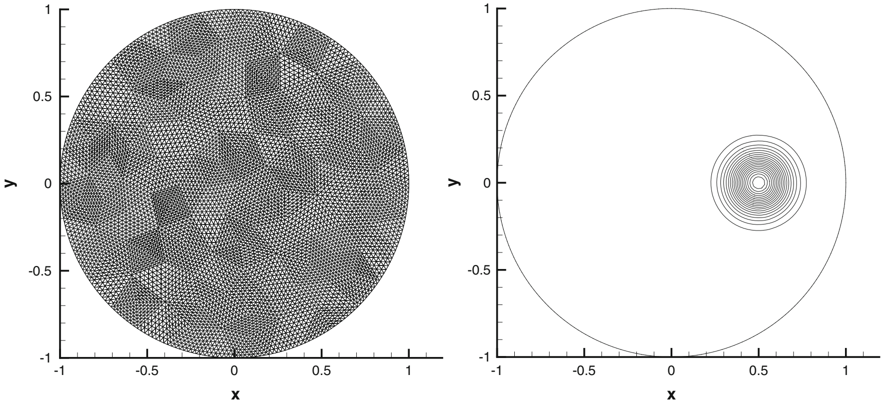


Fig. 1 Exact solution after n rotations ($n \in \mathbb{N}$) and plot of the degrees of freedoms

The domain is a circle with center $(0, 0)$ and radius $R = 1$. The mesh representing all the degrees of freedom is displayed in Fig. 1: The quadratic elements have 6 degrees of freedom (the vertices and the mid-points of the edges). These degrees of freedom are also used for the linear element just by mesh refinement. There are 7047 degrees of freedom here, so $h \approx \sqrt{\frac{\pi}{7047}} \approx 0.021$ which is relatively coarse. On the same figure, we represent the exact solution. The time step is evaluated as the minimum of the Δt_K defined by:

$$\Delta t_K = \text{CFL}\ \frac{h_K}{||\overline{\mathbf{a}}_K||}$$

where h_K is the length of the smallest edge of K and $\overline{\mathbf{a}}_K$ is the speed at the centroid. Since the elements for the $\mathbb{P}^1$ simulations are obtained from those of the $\mathbb{B}^2$ simulation by splitting, the parameter h_K, for the $\mathbb{P}^1$ simulations, is half of the one for the $\mathbf{B}^2$ simulations. For that reason, the CFL number for the quadratic approximation is half of the one chosen for the linear simulations, namely 0.6 instead of 0.3: we run with the same time step. By the way, we have not yet conducted a rigorous study of the CFL condition, but all experiments indicate that the quadratic simulations can be safely run with $CFL = 0.5$.

Figure 2 displays the results for the $\mathbb{P}^1$ approximation, while Fig. 3 shows those obtained for the quadratic approximation. The baseline schemes are the SUPG and the Galerkin scheme with jumps.

In Fig. 2, the same isolines are represented for the three results. We can see that after 10 rotations, the results of the Galerkin+jump scheme look pretty good despite the coarse resolution. The minimum and maximum are -0.012 and 0.762. For the SUPG results, after 1 rotation, the minimum/maximum are -0.004 and 1.02. After 2 rotations we have -0.047 and 1.02. This is better that what is obtained for Fig. 2c, but the dispersive effects are much more important for the SUPG scheme as it can

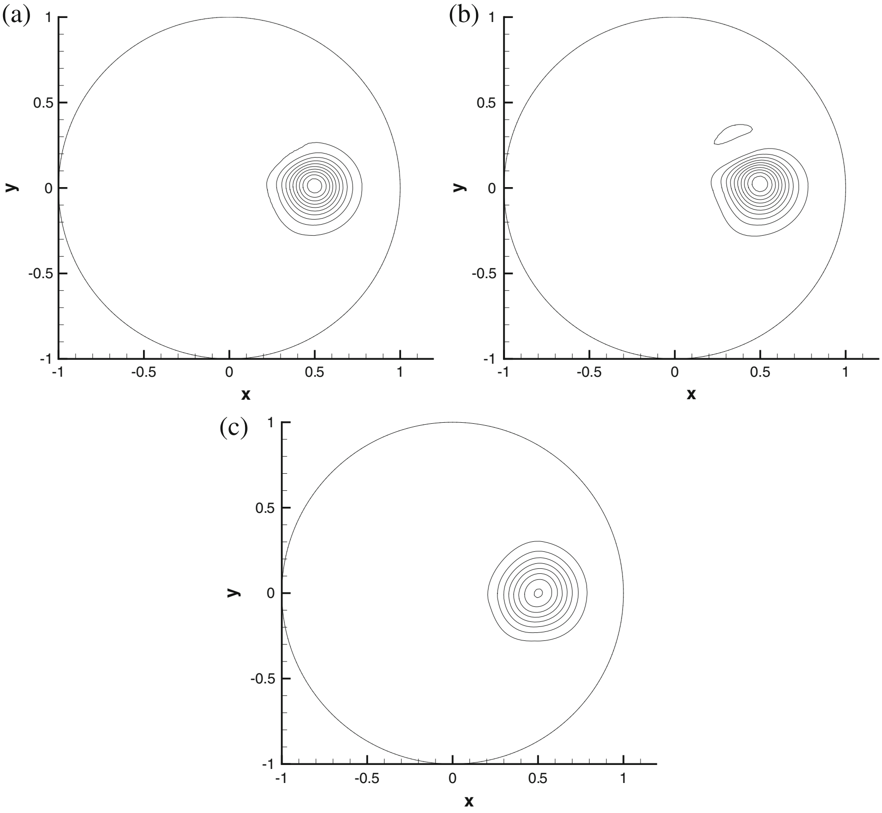


Fig. 2 Results for the $\mathbb{P}^1$ approximation: (**a**) with SUPG, after 1 rotation, (**b**) with SUPG after 2 rotations, (**c**) with Galerkin+Jump after 10 rotations. The same isolines are represented

be seen on Fig. 2b: this is why we have not shown further results for the SUPG/P1 case.

In Fig. 3, we show similar results obtained with the quadratic approximation. Again, the Galerkin+jump method is way less dispersive that the SUPG (stopped after only one rotation this time). We have found that if we perform 4, 6 or 8 iterations of the defect correction, the quality of the SUPG improves a lot, but the cost becomes prohibitive with respect to the Galerkin+jump method for which, after 10 rotations, the min/max are −0.0044 and 0.95. We also see that the solution improves a lot with respect to linear elements, for example in terms of min/max values. There is however some dispersion, if we compare with the exact solution.

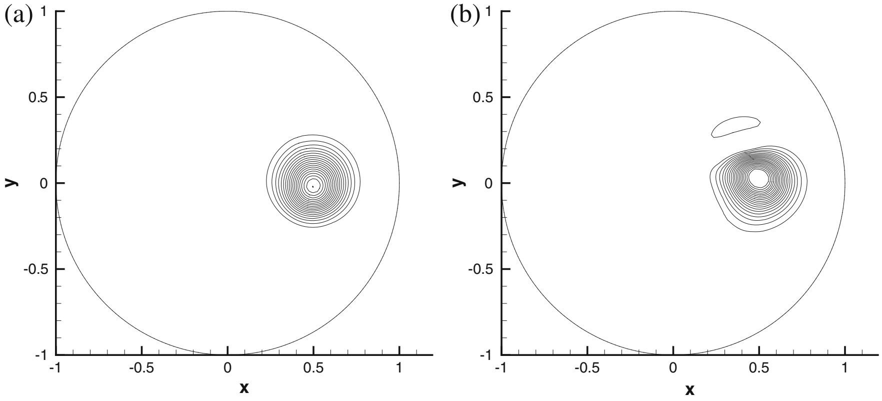


Fig. 3 Results for the $\mathbb{B}^2$ approximation: (**a**) with Galerkin+Jump after 10 rotations, (**b**) with SUPG after 1 rotations, (**c**). The same isolines are represented

4 Conclusions, Perspectives

The paper deals with the numerical approximation of linear scalar hyperbolic problems. We have shown, by carefully choosing the spatial approximation, and by using a non standard time step discretization, that it is possible to avoid the use of mass matrix in this problem, contrarily to what is usually thought about. The cost, on paper, is similar to a standard Runge-Kutta scheme, at least if we consider second and third order in time. In a preliminary work, we have had similar results for the 1D advection problem, which are not shown here. We had also obtained the expected convergence slope.

A lot remains to be done. First, we have found experimental CFL conditions but this has to be rationalized by a numerical analysis. This method needs to be extended to non-linear problems. Preliminary results seems promising, but the results need to be checked on a wider range of problems, this is why we have not reported them here. Last, this method needs to be extended to systems, for example the Euler equations of fluid mechanics.

Acknowledgements This research was funded by SNFS grant # 200021_15360. Early discussions with Mario Ricchiuto (INRIA Bordeaux Sud Ouest, France) are acknowledged.

References

1. R. Abgrall, Some comments about high order approximation of unsteady linear and non linear hyperbolic problems by continuous finite elements (2016, in preparation)
2. R. Abgrall, J. Trefilick, An example of high order residual distribution scheme using non-Lagrange elements. J. Sci. Comput. **45**(1–3), 64–89 (2010)

3. E. Burman, P. Hansbo, Edge stabilization for Galerkin approximation of convection-diffusin-reaction problems. Comput. Methods Appl. Mech. Eng. **193**, 1437–1453 (2004)
4. G. Cohen, P. Joly, J.E. Roberts, N. Tordjman, High order triangular finite elements with mass lumping for the wave equation. SIAM J. Numer. Anal. **38**(6), 2047–2078 (2001)
5. J.A. Cottrell, T.J.R. Hughes, Y. Bazilevs, *Isogeometric Analysis: Toward Integration of CAD and FEA* (Wiley, Hoboken, 2009). ISBN 978-0-470-74873-2
6. A. Dutt, L. Greengard, V. Rokhlin, Spectral deferred correction methods for ordinary differential equations. BIT Numer. Math. **40**(2), 241–266 (2000)
7. T.J.R. Hughes, M. Mallet, A new finite element formulation for computational fluid dynamics III. The generalized streamline operator for multi-dimensional advective-diffusive systems. Comput. Methods Appl. Mech. Eng. **58**, 305–328 (1987)
8. C. Johnson, U. Nävert, J. Pitkäranta, Finite element methods for linear hyperbolic problems. Comput. Methods Appl. Mech. Eng. **45**, 285–312 (1984)
9. S. Jund, S. Salmon, Arbitrary high order finite element schemes and high order mass lumping. Int. J. Appl. Math. Comput. Sci. **17**(3), 375–393 (2007)
10. M. Ricchiuto, R. Abgrall, Explicit runge-kutta residual-distribution schemes for time dependent problems. J. Comput. Phys. **229**(16), 5653–5691 (2010)

Two-Dimensional H(*div*)-Conforming Finite Element Spaces with *hp*-Adaptivity

Philippe R.B. Devloo, Agnaldo M. Farias, Sônia M. Gomes, and Denise de Siqueira

Abstract The purpose of the paper is to analyse the effect of *hp* mesh adaptation when discretized versions of finite element mixed formulations are applied to elliptic problems with singular solutions. Two stable configurations of approximation spaces, based on affine triangular and quadrilateral meshes, are considered for primal and dual (flux) variables. When computing sufficiently smooth solutions using regular meshes, the first configuration gives optimal convergence rates of identical approximation orders for both variables, as well as for the divergence of the flux. For the second configuration, higher convergence rates are obtained for the primal variable. Furthermore, after static condensation is applied, the condensed systems to be solved have the same dimension in both configuration cases, which is proportional to their border flux dimensions. A test problem with a steep interior layer is simulated, and the results demonstrate exponential rates of convergence. Comparison of the results obtained with H^1-conforming formulation are also presented.

1 Introduction

Several methods have been developed for the construction of $\mathbf{H}(div)$-conforming approximation spaces to be applied in flux approximations of the mixed finite element formulation. In some contexts the vector basis functions are constructed on the master element, which is mapped to the elements of the partition using Piola

P.R.B. Devloo
FEC – Universidade Estadual de Campinas, Campinas SP, Brazil
e-mail: phil@fec.unicamp.br

A.M. Farias • S.M. Gomes (✉)
IMECC – Universidade Estadual de Campinas, Campinas SP, Brazil
e-mail: agnaldofarias.mg@gmail.com; soniag@ime.unicamp.br

D.-de Siqueira
Departamento de Matemática, UTFPR 80230-901, Curitiba, PR, Brazil
e-mail: denisesiqueira@utfpr.edu.br

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_9

transformations, as described in [1, 2, 8]. The constructions of hierarchical high order spaces in [5, 7, 10, 11] are based on the properties of the De Rham complex.

Another methodology is proposed in [9] for the construction of hierarchical high order $\mathbf{H}(div)$-conforming approximation spaces based on affine triangular or quadrilateral elements, which has been extended to *hp*-adaptive meshes in [6], and to three-dimensional affine tetrahedral, hexahedral and prismatic meshes in [4]. The principle is to choose appropriate constant vector fields, based on the geometry of each element, which are multiplied by an available set of H^1 hierarchical scalar basis functions to form vectorial shape functions. The assemblage of them, having the characteristic property of $\mathbf{H}(div)$-conforming functions of continuous normal components over element interfaces, is a direct consequence of the properties of the properly chosen vector fields, and of the continuity of the scalar basis functions.

As described in [4], these vectorial shape functions can be combined in different ways to form $\mathbf{H}(div)$-conforming approximation spaces to be applied for flux approximations in discretized versions of the mixed formulation for elliptic problems. In all configurations, the divergence of the dual space and the primal approximation space coincide. There is a first configuration that gives optimal convergence rates of identical approximation orders for primal and dual (flux) variables, as well as for the divergence of the flux, when computing sufficiently smooth solutions using regular meshes. For a second configuration, the accuracy of the primal variable can be enhanced by increasing its approximation order and by enriching the dual space with some properly chosen internal shape functions. Using static condensation, the global condensed matrices to be solved in these two types of space configuration have the same dimension, which is proportional to the dimension of border fluxes.

The purpose of the present paper is to analyse the effect of *hp* mesh adaptation on these space configurations when applied to singular problems. A test problem with a steep interior layer is simulated, and the results demonstrate exponential rates of convergence. Comparison of the results obtained with H^1-conforming formulation are also presented. The implementations are performed in the NeoPZ [1] computational platform, which is an open-source object-oriented project providing a comprehensive set of high performance tools for finite element simulations, including *hp* adaptivity [3].

2 Approximation Spaces in $\mathbf{H}(div, \Omega)$

Let Γ be a mesh on a domain $\Omega \subset \mathbb{R}^2$ formed by elements K. The approximation subspaces in

$$\mathbf{H}(div, \Omega) = \left\{\mathbf{q} \in L^2(\Omega) \times L^2(\Omega); \nabla.\mathbf{q} \in L^2(\Omega)\right\},$$

[1] http://github.com/labmec/neopz

which are defined piecewise over the elements of Γ, require that the local pieces $\mathbf{q}_K = \mathbf{q}|_K$ should be assembled by keeping continuous normal components across common element edges. We shall be concerned with affine triangular or quadrilateral meshes, without any limitation on hanging sides, and varying approximation order distribution $\mathbf{k} = (k_K)$. The proposed methodology used for the construction of such kind of approximation subspaces follows a sequence of steps described below. For more details, we refer to [9], in the case of regular meshes, and to [6] for the case of *hp*-adaptive meshes.

1. For each element K, there is an affine geometric mapping $\mathbf{x} : \hat{K} \to K$, associating each point $\boldsymbol{\xi} \in \hat{K}$ of the (rectangular or triangular) master element $\hat{K}$ to a point $\mathbf{p} = \mathbf{x}(\boldsymbol{\xi}) \in K$.
2. A family of hierarchical bases $\mathbf{B}^K_{k_K} = \{\boldsymbol{\Phi}\}$ is given, where the parameter k_K refers to the degree of the polynomials in $\mathscr{P}_{k_K}$ used in their definitions (of maximum degree, for quadrilateral elements, or of total degree, for triangular elements), as proposed in [9]. The principle is to choose appropriate constant vector fields $\mathbf{v}$, based on the geometry of the element, which are multiplied by an available set of H^1 hierarchical scalar basis functions φ to form a vectorial shape function $\boldsymbol{\Phi} = \varphi\hat{\mathbf{v}}$. There are shape functions of interior type, with vanishing normal components over all element edges. Otherwise, $\boldsymbol{\Phi}$ is classified as of edge type, and its normal component on the edge associated to it coincides with the restriction of the scalar shape function φ used in its definition, and vanishes over the other edges.
3. Construction of approximation subspaces of $\mathbf{H}(\text{div}, \Omega)$ formed by functions $\mathbf{q} \in \left[L^2(\Omega)\right]^2$, which are defined piecewise over the elements of Γ by local functions $\mathbf{q}_K = \mathbf{q}|_K \in span\ \mathbf{B}^K_{k_K} \subset \mathbf{H}(\text{div}, K)$. As described in [6], the pieces can be easily assembled to get continuous normal components on the elements interfaces. This property is obtained as a consequence of the particular properties satisfied by the proposed vectorial shape functions, and the continuity of the scalar shape functions used in their construction.

3 Application to Mixed Finite Element Formulation

Given $f \in L^2(\Omega)$, boundary values u_D and g for Dirichlet and Neumann conditions enforced on $\partial\Omega_D$ and $\partial\Omega_N$, consider the variational mixed formulation of finding $u \in L^2(\Omega)$ and $\boldsymbol{\sigma} \in \mathbf{V} = \{\mathbf{q} \in \mathbf{H}(div, \Omega); \boldsymbol{\sigma} \cdot \boldsymbol{\eta}|_{\partial\Omega_N} = -g\}$, such that, for all $v \in L^2(\Omega)$, and $\mathbf{q} \in \mathbf{V}_0 = \{\mathbf{q} \in \mathbf{H}(div, \Omega); \mathbf{q} \cdot \boldsymbol{\eta}|_{\partial\Omega_N} = 0\}$,

$$\int_\Omega \boldsymbol{\sigma} \cdot \mathbf{q}\, d\Omega - \int_\Omega u\, \nabla \cdot \mathbf{q}\, d\Omega = -\int_{\partial\Omega_D} u_D\, \mathbf{q} \cdot \boldsymbol{\eta}\, ds,$$

$$-\int_\Omega \nabla \cdot \boldsymbol{\sigma}\ v\, d\Omega = -\int_\Omega f v\, d\Omega.$$

Approximation Spaces Following the developments in [4], we shall consider two stable configuration cases for approximation spaces to be used for primal u and dual $\boldsymbol{\sigma}$ variables in discretized versions of the mixed formulation. In both cases, the primal variable is approximated in subspaces of $L^2(\Omega)$ formed by piecewise functions $u|_K = u_K$, without any continuity constraint, as in typical discretized mixed formulations [2].

The first configuration considers polynomials $u_K \in \mathscr{P}_{k_K}$, and the dual variable $\boldsymbol{\sigma}$ is sought in approximation spaces $\subset \mathbf{H}(div, \Omega)$ formed by vectorial functions $\mathbf{q}$ such that $\boldsymbol{q}_K = \boldsymbol{q}|_K \in span\ \mathbf{B}^{K*}_{k_K}$, where the bases $\mathbf{B}^{K*}_{k_K} \subset \mathbf{B}^{K}_{k_K+1}$ are formed by enriching $\mathbf{B}^{K}_{k_K}$ with interior shape functions $\boldsymbol{\Phi} \in \mathbf{B}^{K}_{k_K+1}$ whose divergence $\nabla \cdot \boldsymbol{\Phi} \in \mathscr{P}_{k_K}$. The resulting set of approximations spaces is classified as being of $\mathbf{P}^*_\mathbf{k}$ $P_\mathbf{k}$ type.

Another type of approximation configuration is classified as being of $\mathbf{P}^{**}_\mathbf{k}$ $P_{\mathbf{k}+1}$ type, where the primal approximations u_K are in $\mathscr{P}_{k_K+1}$, and $\mathbf{P}^{**}_k$ refers to vectorial approximation spaces spanned by bases $\mathbf{B}^{K**}_{k_K} \subset \mathbf{B}^{K*}_{k_K+1}$, where the edge functions are restricted to those ones of $\mathbf{P}^*_\mathbf{k}$ type.

As explained in [4], when computing sufficiently smooth solutions using $\mathbf{P}^*_\mathbf{k}$ $P_\mathbf{k}$ space configurations based on affine regular meshes, optimal convergence rates of identical approximation orders $k + 1$ are obtained for primal and dual variables, as well for $\nabla \cdot \boldsymbol{\sigma}$. For the $\mathbf{P}^{**}_\mathbf{k}$ $P_{\mathbf{k}+1}$ configuration, higher convergence rate of order $k + 2$ is obtained for the primal variable. Furthermore, after static condensation is applied, the condensed systems to be solved only involve the flux edge terms and a constant value for u in each element, and thus they have the same dimension in both configuration cases.

Test Problem The problem is defined over the domain of $\Omega = [0, 1] \times [0, 1]$, and the load function f is chosen such that the model problem has exact solution given by

$$u(x, y) = \frac{\pi}{2} - \arctan\left[\alpha\left(\sqrt{(x - 1.25)^2 + (y + 0.25)^2} - \frac{\pi}{3}\right)\right],$$

having strong gradients with magnitude determined by the parameter $\alpha = 200$ in the proximity to the circumference centred at the point (1.25, –0.25), with radius $\pi/3$. Plots of the exact solution $u(x, y)$ and its gradient magnitude are presented in Fig. 1.

Adaptive *hp*-Refinement Process We consider a sequence of *hp*-adaptive meshes with either quadrilateral or triangular geometries, with variable polynomial degree distributions. To construct them, firstly, split the domain into two regions: the region near the singularity and the smooth part, elsewhere. In the region where the solution is smooth, p refinement is adopted in order to produce exponential convergence rates there. In the central region, *hp* refinement is employed in order to generate approximation spaces which better capture the singular behaviour. The initial mesh is composed of uniform elements with mesh size 2^{-3}, and $p = 2$ in the smooth part, and mesh size 2^{-4} and $p = 3$ in the region of the singularity. Then, the refinement process follows a sequence of steps $\ell = 2, 3$, and 4 by first increasing by 1 the

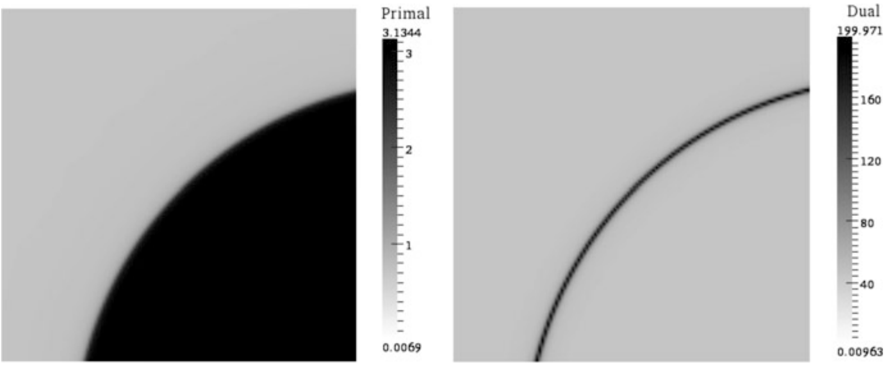


Fig. 1 Exact solution: primal (*left side*) and dual (*right side*) variables

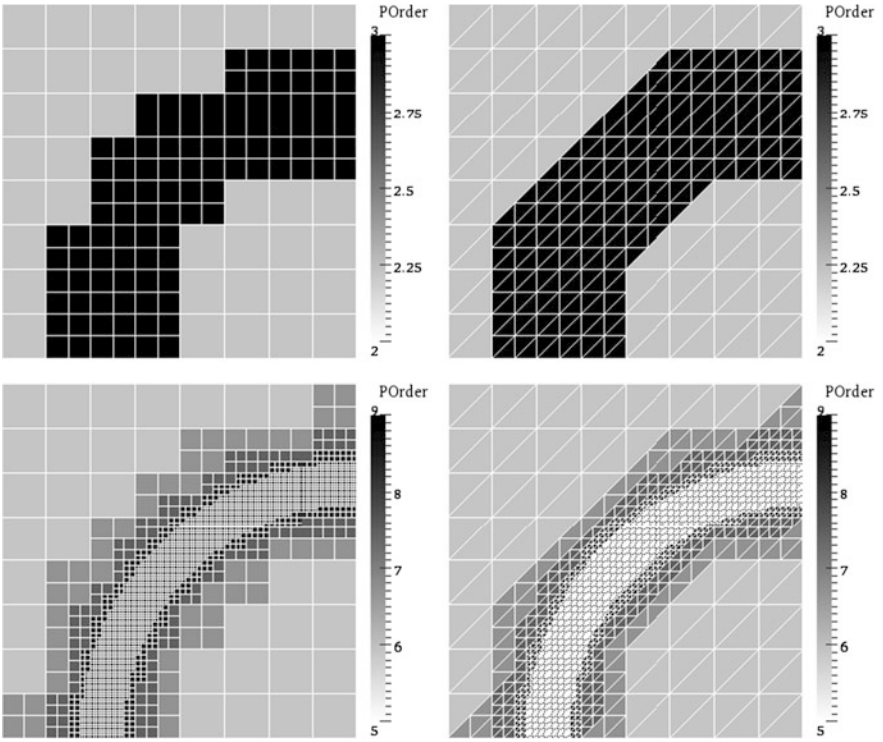


Fig. 2 Illustration of the *hp* refinement process: initial mesh (*top-side*), and mesh at the final refinement step (*bottom-side*) for quadrilateral (*left-side*) and triangular (*right-side*) geometries

approximation order of all elements of the previous step, and then by subdividing the elements intersecting a layer of diameter $2^{-\ell}$ around the singularity curve, and by further increasing their approximation order by 1. Figure 2 illustrates the *hp* refinement process at the initial step, and at the final refinement level.

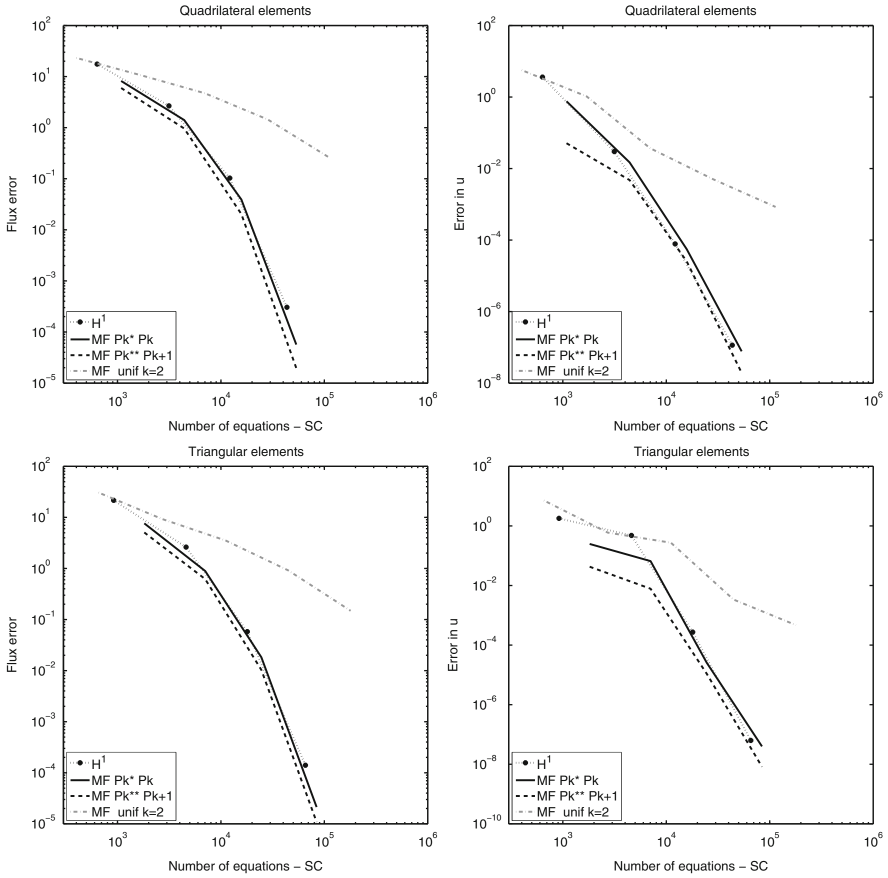


Fig. 3 L^2-error curves in terms of the number of degrees of freedom for the dual (*left side*) and primal (*right side*) variables using mixed formulation and approximation spaces of type $\mathbf{P}_\mathbf{k}^* \ P_\mathbf{k}$ (*continuous curves*), and $\mathbf{P}_\mathbf{k}^{**} \ P_{\mathbf{k}+1}$ (*dashed curves*) based on *hp*-meshes with quadrilateral (*top-side*) and triangular (*bottom-side*) geometries. The *dashed-dotted curves* correspond to simulations for uniform meshes with $\mathbf{P}_\mathbf{2}^* \ P_\mathbf{2}$ configuration. For comparison, results for H^1-conforming formulation based on the same *hp*-meshes are also included (*dotted curves*)

Our purpose is to use these kinds of meshes for the simulation of the test problem by the mixed formulations using the space configurations of $\mathbf{P}_\mathbf{k}^* \ P_\mathbf{k}$ and $\mathbf{P}_\mathbf{k}^{**} \ P_{\mathbf{k}+1}$ types. As expected, the application of *hp* refinement to the singular problem improves considerably the performance of the methods, with exponential rates of convergence. Furthermore, the accuracy in the primal variable improves when $\mathbf{P}_\mathbf{k}^{**} \ P_{\mathbf{k}+1}$ configuration is applied in the mixed formulation. Figure 3 shows the calculated L^2-norms of the dual $\boldsymbol{\sigma}$ and primal u errors using these sequences of *hp*-adaptive meshes versus the number of equations solved after static condensation. For comparison, results for the H^1-conforming formulation based on

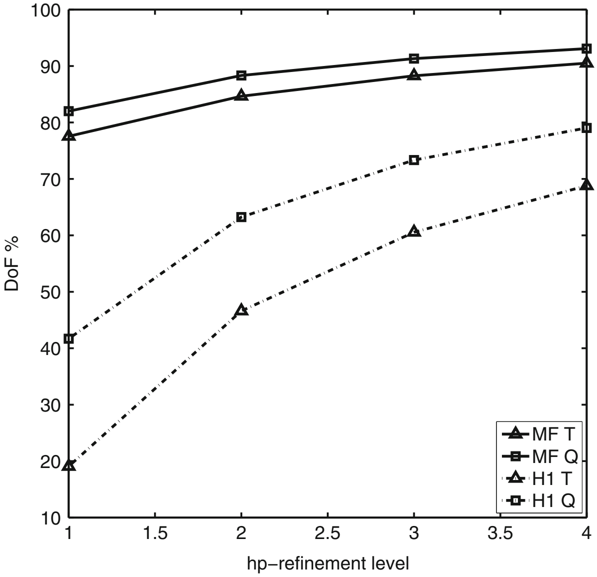


Fig. 4 Percentage of condensed degrees of freedom in the mixed method using the $P_k^* P_k$ space configuration (*continuous lines*) and in the H^1-conforming method (*dashed-dotted lines*), applied to quadrilateral and triangular *hp*-meshes

the same *hp*-meshes, and for the mixed formulation with uniform meshes and $\mathbf{P}_2^*$ P_2 configuration are plotted. For the experiments with H^1 conforming approximations, the performance in terms of accuracy versus degrees of freedom is similar to the experiments with the mixed formulation.

The effect of static condensation is also verified in terms of the size reduction of the global system to be solved, which is more significant in the mixed formulation, with increasing order of approximation, and with quadrilateral meshes, as compared with triangular ones. At the finest levels of mesh refinement, the number of condensed equations in the mixed formulation amounts to more than 90 %, as shown in Fig. 4, meaning that the size of the condensed system to be solved is less than 10 % of the total number of equations. This fact demonstrates the potential benefit of using **H**(*div*) approximation spaces in parallel computers.

Acknowledgements R. B. Devloo and S. M. Gomes thankfully acknowledge financial support from CNPq – the Brazilian Research Council. The work of A. M. Farias is supported by a posdoctoral grant by CAPES Foundation, within the Ministry of Education in Brazil. The authors thank T. L. D. Forti for given the adaptive *hp* meshes, and to P. C. A. Lucci for helping with the graphics.

References

1. F. Brezzi, J. Douglas, L.D. Marini, Two families of mixed finite elements for second order elliptic problems. Numer. Math. **47**, 217–235 (1985)
2. F. Brezzi, M. Fortin, *Mixed and Hybrid Finite Element Methods*. Springer Series in Computational Mathematics, vol. 15 (Springer, NewYork, 1991)
3. J.L.D. Calle, P.R.B. Devloo, S.M. Gomes, Implementation of continuous *hp*-adaptive finite element spaces without limitations on hanging sides and distribution of approximation orders. Comput. Math. Appl. **70**(5), 1051–1069 (2015)
4. D.A. Castro, P.R.B. Devloo, A.M. Farias, S.M. Gomes, D. Siqueira, Three dimensional hierarchical mixed finite element approximations with enhanced primal variable accuracy. Comput. Methods Appl. Mech. Eng. **306**, 479–502 (2016)
5. L. Demkowicz, *Polynomial Exact Sequence and Projection-Based Interpolation with Application Maxwell Equations*, ed. by D. Boffi, L. Gastaldi. Mixed Finite Elements, Compatibility Conditions and Applications, Lecture Notes in Mathematics, vol. 1939 (Springer, Heidelberg, 2007), pp. 101–158
6. P.R.B. Devloo, A.M. Farias, S.M. Gomes, D. Siqueira, Two-dimensional *hp*-adaptive finite element spaces for mixed formulations. Math. Comput. Simul. **126**,104–122 (2016)
7. F. Fuentes, B. Keith, L. Demkowicz, S. Nagaraj, Orientation embedded high order shape functions for the exact sequence elements of all shapes. Comput. Math. Appl. **70**(4), 353–458 (2015)
8. J.C. Nedelec, A new family of mixed finite elements in $\mathbb{R}^3$. Numer. Math. **50**, 57–81 (1986)
9. D. Siqueira, P.R.B. Devloo, S.M. Gomes, A new procedure for the construction of hierarchical high order H(div) and H(curl) finite element spaces. J. Comput. Appl. Math. **240**, 204–214 (2013)
10. P. Solin, K. Segeth, I. Dolezel, *Higher-Order Finite Element Methods* (Chapman – Hall/CRC, Boca Raton, 2004)
11. S. Zaglamayr, Hight order finite element methods for electromagnetic field computation. Ph.D. thesis, Johannes Keppler Universität Linz, 2006

Finite Elements for the Navier-Stokes Problem with Outflow Condition

Daniel Arndt, Malte Braack, and Gert Lube

Abstract This work is devoted to the Directional Do-Nothing (DDN) condition as an outflow boundary condition for the incompressible Navier-Stokes equation. In contrast to the Classical Do-Nothing (CDN) condition, we have stability, existence of weak solutions and, in the case of small data, also uniqueness. We derive an a priori error estimate for this outflow condition for finite element discretizations with inf-sup stable pairs. Stabilization terms account for dominant convection and the divergence free constraint. Numerical examples demonstrate the stability of the method.

1 Introduction

The classical do-nothing condition is very often prescribed at outflow boundaries for fluid dynamical problems. However, in the case of the Navier-Stokes equations in a domain $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$, not even existence of weak solutions can be shown, see [10]. The reason is that this boundary condition does not exhibit any control about inflow across such boundaries, see [4]. This has also severe impact onto the stability of numerical algorithms for flows at higher Reynolds numbers. Denoting the velocity field by $\boldsymbol{u}$ and the pressure by p, the directional do-nothing (DDN) boundary condition

$$\nu \nabla \boldsymbol{u} \cdot \boldsymbol{n} - p\boldsymbol{n} - \beta(\boldsymbol{u} \cdot \boldsymbol{n})_- \boldsymbol{u} = 0 \quad \text{at } S_1 \tag{1}$$

on $S_1 \subseteq \partial\Omega$ with normal vector $\boldsymbol{n}$ and a parameter $\beta \geq 0$ is one possibility to circumvent this disadvantage. Here, $(\boldsymbol{u} \cdot \boldsymbol{n})_-(x) = \min(0, \boldsymbol{u}(x) \cdot \boldsymbol{n}(x))$ denotes the negative part of the flux across the boundary at $x \in \partial\Omega$. In particular, existence of

D. Arndt • G. Lube
University of Göttingen, Göttingen, Germany
e-mail: d.arndt@math.uni-goettingen.de; lube@math.uni-goettingen.de

M. Braack (✉)
Mathematical Seminar, University of Kiel, Kiel, Germany
e-mail: braack@math.uni-kiel.de

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_10

weak solutions is proved in [4], and in several applications the stability is enhanced compared to the classical do-nothing condition ($\beta = 0$), see e.g. [2, 12].

In the case of pure outflow, i.e. if $\boldsymbol{u}\cdot\boldsymbol{n} \geq 0$ on S_1, this condition is identical to the classical do-nothing condition (CDN). In particular, it reproduces Poiseuille flow for a laminar flow in a tube with parabolic inflow.

The outflow condition (1) has similarities with the *convective boundary condition* in [5, 6], but no reference solution nor Stokes solution on a larger domain is needed. We also like to refer to the recent work [8] where a different open boundary condition is proposed which makes use of a smoothed step function and overcomes backflow instabilities as well.

2 Variational Formulation

The variational spaces for velocity and pressure are given by

$$\begin{aligned} \boldsymbol{V} &:= \{\boldsymbol{u} \in H^1(\Omega)^d \,|\, \boldsymbol{u} = 0 \text{ a.e. on } S_0\}, \\ Q &:= L^2(\Omega), \end{aligned}$$

respectively. The norm in $L^2(\omega)$ (and in $L^2(\omega)^d$) for $\omega \subseteq \Omega$ is denoted by $\|\cdot\|_\omega$. For $\omega = \Omega$ we surpress the index. The $H^{-1}(\Omega)$-norm is denoted by $\|\cdot\|_{-1}$. In order to formulate the Navier-Stokes system in variational form we consider the decomposition

$$((\boldsymbol{w}\cdot\nabla)\boldsymbol{u}, \boldsymbol{\phi}) = \frac{1}{2}\left(((\boldsymbol{w}\cdot\nabla)\boldsymbol{u}, \boldsymbol{\phi}) - (\boldsymbol{u}, (\boldsymbol{w}\cdot\nabla)\boldsymbol{\phi})\right) + \frac{1}{2}\int_{\partial\Omega} (\boldsymbol{w}\cdot\boldsymbol{n})\boldsymbol{u}\cdot\boldsymbol{\phi}\, ds,$$

of the convective term for divergence free vector fields $\boldsymbol{w}$ and use the notation

$$\begin{aligned} c(\boldsymbol{w}; \boldsymbol{u}, \boldsymbol{\phi}) := \tfrac{1}{2}((\boldsymbol{w}\cdot\nabla)\boldsymbol{u}, \boldsymbol{\phi}) - \tfrac{1}{2}(\boldsymbol{u}, (\boldsymbol{w}\cdot\nabla)\boldsymbol{\phi}) \\ + \int_{S_1} \left(\tfrac{1}{2}(\boldsymbol{w}\cdot\boldsymbol{n}) - \beta(\boldsymbol{w}\cdot\boldsymbol{n})_-\right)\boldsymbol{u}\cdot\boldsymbol{\phi}\, ds. \end{aligned}$$

Lemma 1 *The nonlinear convective term can be expressed by*

$$c(\boldsymbol{w}; \boldsymbol{u}, \boldsymbol{\phi}) = ((\boldsymbol{w}\cdot\nabla)\boldsymbol{u}, \boldsymbol{\phi}) + \tfrac{1}{2}(div\, \boldsymbol{w}\, \boldsymbol{u}, \boldsymbol{\phi}) - \beta\int_{S_1} (\boldsymbol{w}\cdot\boldsymbol{n})_- \boldsymbol{u}\cdot\boldsymbol{\phi}\, ds$$

for all $\boldsymbol{w}, \boldsymbol{u}, \boldsymbol{\phi} \in \boldsymbol{V}$.

Proof This identity follows easily by integration by parts.

The semi-linear form for the Navier-Stokes system with DDN condition reads

$$A(\boldsymbol{w}; \boldsymbol{u}, p; \boldsymbol{\phi}, \chi) := c(\boldsymbol{w}; \boldsymbol{u}, \boldsymbol{\phi}) + (\nu\nabla\boldsymbol{u}, \nabla\boldsymbol{\phi}) - (p, \operatorname{div}\boldsymbol{\phi}) + (\operatorname{div}\boldsymbol{u}, \chi).$$

We seek $(\boldsymbol{u}, p) \in \boldsymbol{V} \times Q$ s.t.

$$(\partial_t \boldsymbol{u}, \boldsymbol{\phi}) + A(\boldsymbol{u}; \boldsymbol{u}, p; \boldsymbol{\phi}, \chi) = \langle \boldsymbol{f}, \boldsymbol{\phi} \rangle \qquad \forall \boldsymbol{\phi} \in \boldsymbol{V}, \forall \chi \in Q. \tag{2}$$

Diagonal testing with the solution, $\boldsymbol{\phi} := \boldsymbol{u}$, $\chi := p$, results in

$$\frac{1}{2}\partial_t \|\boldsymbol{u}\|^2 + \nu \|\nabla \boldsymbol{u}\|^2 + \int_{S_1} \left(\frac{1}{2}(\boldsymbol{u} \cdot \boldsymbol{n})_+ + \left(\frac{1}{2} - \beta \right) (\boldsymbol{u} \cdot \boldsymbol{n})_- \right) |\boldsymbol{u}|^2 ds \\ \leq \|\boldsymbol{f}\|_{-1} \|\nabla \boldsymbol{u}\|.$$

For $\beta \geq 1/2$ the arising boundary integral is non-negative. This property is needed to show existence of weak solutions, see techniques in [4]. Moreover, the solution is unique in the case of small data, see [3].

3 Finite Element Discretization

For the discretization of (2) in space we use inf-sup stable finite elements of order k for $\boldsymbol{u}_h$, for instance the classical Taylor-Hood element Q_2/Q_1 on quadrilaterals (for $d = 2$). Due to the fact that we use a divergence-free projection in the analysis below, we require for $d = 3$ on hexahedrons Q_3/Q_2 elements, see [9]. It is well-known that the convective terms and the divergence-free constraint should be stabilized in order to obtain more accurate discrete solutions with enhanced divergence properties and less over- and undershoots. We use a combination of div-div stabilization and local projection (LPS) of the convective terms

$$S_h(\boldsymbol{w}; \boldsymbol{u}, \boldsymbol{\phi}) := \sum_{M \in \mathscr{M}_h} \gamma_M (\operatorname{div} \boldsymbol{u}, \operatorname{div} \boldsymbol{\phi})_M + \alpha_M (\kappa_M[(\boldsymbol{w}_M \cdot \nabla)\boldsymbol{u}], \kappa_M[(\boldsymbol{w}_M \cdot \nabla)\boldsymbol{\phi}])_M$$

with local fluctuation operator $\kappa_M : L^2(M) \to L^2(M)$ on patches M, and piecewise constant approximation $\boldsymbol{w}_M$ of $\boldsymbol{w}$. We allow for the one-level ($M \in \mathscr{T}_h$) or the two-level ($M \in \mathscr{T}_{2h}$) variant, but the common requirements according to [11] should be satisfied. The stabilization parameter γ_M for the divergence stabilization is patch-wise constant in the following range:

$$0 < \gamma_0 h_{max} \leq \gamma_M \ \leq \ \gamma_{max}, \tag{3}$$

with positive constants $\gamma_0, \gamma_{max} > 0$ and the maximal mesh size $h_{max} = \max\{h_T : T \in \mathscr{T}_h\}$. The LPS parameter α_M must be non-negative (may vanish) and may depend on $\boldsymbol{u}_M$ but is bounded ($\alpha_0 \geq 0$):

$$0 \leq \alpha_M \ \leq \ \alpha_0 |\boldsymbol{u}_M|^{-2}. \tag{4}$$

Similar to the work [1] we assume for the a priori estimate the following local approximation property:

$$\|\boldsymbol{u} - \boldsymbol{u}_M\|_{L^\infty(M)} \leq C\|\boldsymbol{u}\|_{L^\infty(M)}. \tag{5}$$

The semi-discrete system consists in seeking $\boldsymbol{u}_h \in \boldsymbol{V}_h, p_h \in Q_h$ s.t.

$$(\partial_t \boldsymbol{u}_h, \boldsymbol{\phi}) + A(\boldsymbol{u}_h; \boldsymbol{u}_h, p_h; \boldsymbol{\phi}, \chi) + S_h(\boldsymbol{u}_h; \boldsymbol{u}_h, \boldsymbol{\phi}) = \langle \boldsymbol{f}, \boldsymbol{\phi} \rangle \tag{6}$$

for all $\boldsymbol{\phi} \in \boldsymbol{V}_h$ and all $\chi \in Q_h$.

4 A Priori Estimate

For the a priori estimate we split the error $\boldsymbol{e}_u := \boldsymbol{u} - \boldsymbol{u}_h$ into interpolation error $\boldsymbol{\eta}_u := \boldsymbol{u} - \boldsymbol{i}_h \boldsymbol{u}$ and projection error $\boldsymbol{\xi}_u := \boldsymbol{i}_h \boldsymbol{u} - \boldsymbol{u}_h$. Here, $\boldsymbol{i}_h : \boldsymbol{V} \to \boldsymbol{V}_h$ is a divergence-free projection. We use the following norm in $\boldsymbol{V}$:

$$\begin{aligned} |||\boldsymbol{u}|||_{\boldsymbol{u}_h}^2 = \nu\|\nabla \boldsymbol{u}\|^2 + \int_{S_1} \left(\tfrac{1}{2}(\boldsymbol{u}_h \cdot \boldsymbol{n})_+ + (\tfrac{1}{2} - \beta)(\boldsymbol{u}_h \cdot \boldsymbol{n})_-\right) |\boldsymbol{u}|^2 d\sigma \\ + S_h(\boldsymbol{u}_h; \boldsymbol{u}, \boldsymbol{u}). \end{aligned}$$

A bound on the interpolation error $\boldsymbol{\eta}_u$ is well-known, see [9]. Therefore we focus on the projection error.

Theorem 2 *Under the previous assumptions* (3), (4) *and* (5), *enough regularity of the continuous solution* $\boldsymbol{u}$, p, *and* $\beta > 1/2$ *it holds for the projection error:*

$$\|\boldsymbol{\xi}_u\|_{L^\infty(0,T;L^2(\Omega))}^2 + \int_0^T |||\boldsymbol{\xi}_u(t)|||_{\boldsymbol{u}_h}^2 dt \leq C \int_0^T e^{C_G(t-\tau)} \sum_M \phi_M(\tau) d\tau$$

with the Gronwall constant

$$C_G := c(1 + |\boldsymbol{u}|_{L^\infty(0,T;W^{1,\infty}(\Omega))} + h\|\boldsymbol{u}\|_{L^\infty(0,T;W^{1,\infty}(\Omega))}^2 + (1 + \nu^{-1})\|\boldsymbol{u}\|_{L^\infty(S_1)}),$$

and the quantity ϕ_M *depending on* $\boldsymbol{u}$ *and on the interpolation errors* $\boldsymbol{\eta}_u$, η_p*:*

$$\phi_M := \|\partial_t \boldsymbol{\eta}_u\|_M^2 + (c_1 + c_3)\|\nabla \boldsymbol{\eta}_u\|_M^2 + c_2\|\boldsymbol{\eta}_u\|_M^2 + c_3\|\kappa_M(\nabla \boldsymbol{u})\|_M^2 + c_4\|\eta_p\|_M^2,$$

and coefficients $c_1, \ldots, c_4$*:*

$$c_1 = \nu + \gamma_M,\ c_2 = h_M^{-2} + \nu^{-1}\|\boldsymbol{u}\|_{L^\infty(M)}^2,\ c_3 = \alpha_M|\boldsymbol{u}|_M^2,\ c_4 = (\nu + \gamma_M)^{-1}.$$

This bound is similar to the one published in [1] (for $S_1 = \emptyset$). The difference is the additional term $\nu^{-1}\|\boldsymbol{u}\|_{L^\infty(S_1)}$ in the Gronwall constant.

Proof In the first step we subtract (2) and (6) and perform diagonal testing. Due to the additive splitting of the error, $\boldsymbol{e}_u = \boldsymbol{\eta}_u + \boldsymbol{\xi}_u$, and after reordering terms we arrive at

$$\begin{aligned}\frac{1}{2}\partial_t\|\boldsymbol{\xi}_u\|^2 + |||\boldsymbol{\xi}_u|||^2_{\boldsymbol{u}_h} = &-(\partial_t\boldsymbol{\eta}_u, \boldsymbol{\xi}_u) - \nu(\nabla\boldsymbol{\eta}_u, \nabla\boldsymbol{\xi}_u) + (\operatorname{div}\boldsymbol{\xi}_u, \eta_p) - (\operatorname{div}\boldsymbol{\eta}_u, \xi_p)\\ &-c(\boldsymbol{u}; \boldsymbol{u}, \boldsymbol{\xi}_u) + c(\boldsymbol{u}_h; \boldsymbol{i}_h\boldsymbol{u}, \boldsymbol{\xi}_u) + S_h(\boldsymbol{u}_h; \boldsymbol{i}_h\boldsymbol{u}, \boldsymbol{\xi}_u)\end{aligned}$$

Using Lemma 1 we obtain

$$\begin{aligned}&-c(\boldsymbol{u}; \boldsymbol{u}, \boldsymbol{\xi}_u) + c(\boldsymbol{u}_h; \boldsymbol{i}_h\boldsymbol{u}, \boldsymbol{\xi}_u)\\ =& -(\boldsymbol{u}\cdot\nabla\boldsymbol{u}, \boldsymbol{\xi}_u) + (\boldsymbol{u}_h\cdot\nabla\boldsymbol{i}_h\boldsymbol{u}, \boldsymbol{\xi}_u) + \tfrac{1}{2}(\operatorname{div}\boldsymbol{u}_h\,\boldsymbol{i}_h\boldsymbol{u}, \boldsymbol{\xi}_u)\\ &+\beta\int_{S_1}\left\{(\boldsymbol{u}\cdot\boldsymbol{n})_-\boldsymbol{u} - (\boldsymbol{u}_h\cdot\boldsymbol{n})_-\boldsymbol{i}_h\boldsymbol{u}\right\}\cdot\boldsymbol{\xi}_u ds\\ =& -(\boldsymbol{e}_u\cdot\nabla\boldsymbol{u}, \boldsymbol{\xi}_u) - (\boldsymbol{u}_h\cdot\nabla\boldsymbol{\eta}_u, \boldsymbol{\xi}_u) + \tfrac{1}{2}(\operatorname{div}\boldsymbol{u}_h\,\boldsymbol{i}_h\boldsymbol{u}, \boldsymbol{\xi}_u)\\ &+\beta\int_{S_1}\left\{(\boldsymbol{u}\cdot\boldsymbol{n})_-\boldsymbol{u} - (\boldsymbol{u}_h\cdot\boldsymbol{n})_-\boldsymbol{i}_h\boldsymbol{u}\right\}\cdot\boldsymbol{\xi}_u ds.\end{aligned}$$

Integration by parts a second time yields

$$(\boldsymbol{u}_h\cdot\nabla\boldsymbol{\eta}_u, \boldsymbol{\xi}_u) = -(\boldsymbol{\eta}_u, \boldsymbol{u}_h\cdot\nabla\boldsymbol{\xi}_u) - (\operatorname{div}\boldsymbol{u}_h\,\boldsymbol{\eta}_u, \boldsymbol{\xi}_u) + \int_{S_1}(\boldsymbol{u}_h\cdot\boldsymbol{n})\boldsymbol{\eta}_u\cdot\boldsymbol{\xi}_u\, ds.$$

We obtain the identity

$$\frac{1}{2}\partial_t\|\boldsymbol{\xi}_u\|^2 + |||\boldsymbol{\xi}_u|||^2_{\boldsymbol{u}_h} = R + T,$$

with volume integrals

$$\begin{aligned}R := &-(\partial_t\boldsymbol{\eta}_u, \boldsymbol{\xi}_u) - \nu(\nabla\boldsymbol{\eta}_u, \nabla\boldsymbol{\xi}_u) + (\operatorname{div}\boldsymbol{\xi}_u, \eta_p) - (\operatorname{div}\boldsymbol{\eta}_u, \xi_p)\\ &-(\boldsymbol{e}_u\cdot\nabla\boldsymbol{u}, \boldsymbol{\xi}_u) + \tfrac{1}{2}(\operatorname{div}\boldsymbol{u}_h\,\boldsymbol{i}_h\boldsymbol{u}, \boldsymbol{\xi}_u) + (\boldsymbol{\eta}_u, \boldsymbol{u}_h\cdot\nabla\boldsymbol{\xi}_u)\\ &+(\operatorname{div}\boldsymbol{u}_h\,\boldsymbol{\eta}_u, \boldsymbol{\xi}_u) + S_h(\boldsymbol{u}_h; \boldsymbol{i}_h\boldsymbol{u}, \boldsymbol{\xi}_u),\end{aligned}$$

and boundary integrals

$$T := \int_{S_1}\left\{\beta(\boldsymbol{u}\cdot\boldsymbol{n})_-\boldsymbol{u} - \beta(\boldsymbol{u}_h\cdot\boldsymbol{n})_-\boldsymbol{i}_h\boldsymbol{u} - (\boldsymbol{u}_h\cdot\boldsymbol{n})\boldsymbol{\eta}_u\right\}\cdot\boldsymbol{\xi}_u ds.$$

For R we may use the result of [1]:

$$R \leq \frac{1}{2}\|\partial_t \boldsymbol{\eta}_u\|^2 + C_G\|\boldsymbol{\xi}_u\|^2 + \frac{1}{4}|||\boldsymbol{\xi}_u|||^2_{\boldsymbol{u}_h} + 2\sum_M \phi_M.$$

The techniques in [1] do not require any further integration by parts. Therefore, the approach for the Dirichlet case without any outflow condition also applies to bound R in our case. The remaining terms of T will be bounded in the sequel. A basic calculus yield

$$T = T_1 + T_2$$

with

$$\begin{aligned} T_1 &:= -\int_{S_1} \{(\boldsymbol{u}_h \cdot \boldsymbol{n})_+ + (1-\beta)(\boldsymbol{u}_h \cdot \boldsymbol{n})_-\}\boldsymbol{\eta}_u \cdot \boldsymbol{\xi}_u \, ds, \\ T_2 &:= \beta \int_{S_1} \{(\boldsymbol{u} \cdot \boldsymbol{n})_- - (\boldsymbol{u}_h \cdot \boldsymbol{n})_-\}\boldsymbol{u} \cdot \boldsymbol{\xi}_u \, ds. \end{aligned}$$

Since the triple-norm includes control on the boundary fluxes, T_1 is bounded by $|||\cdot|||_{\boldsymbol{u}_h}$ provided $\beta > \frac{1}{2}$:

$$\begin{aligned} T_1 &= -\int_{S_1} \{(\boldsymbol{u}_h \cdot \boldsymbol{n})_+ + (1-\beta)(\boldsymbol{u}_h \cdot \boldsymbol{n})_-\}\boldsymbol{\eta}_u \cdot \boldsymbol{\xi}_u \, ds \\ &\leq \max\left(2, \frac{|\beta - 1|}{\beta - \frac{1}{2}}\right) |||\boldsymbol{\eta}_u|||_{\boldsymbol{u}_h} |||\boldsymbol{\xi}_u|||_{\boldsymbol{u}_h}. \end{aligned}$$

T_2 can be bounded by the trace theorem in L^1-norm and the product rule with arbitrary $\epsilon > 0$:

$$\begin{aligned} T_2 &= \beta \int_{S_1} \{(\boldsymbol{u} \cdot \boldsymbol{n})_- - (\boldsymbol{u}_h \cdot \boldsymbol{n})_-\}\boldsymbol{u} \cdot \boldsymbol{\xi}_u \, ds \\ &\leq c_S \|\boldsymbol{u}\|_{L^\infty(S_1)} \| |\boldsymbol{e}_u| |\boldsymbol{\xi}_u| \|_{W^{1,1}(\Omega)} \\ &\leq c_S \|\boldsymbol{u}\|_{L^\infty(S_1)} (\| |\boldsymbol{e}_u| |\boldsymbol{\xi}_u| \|_{L^1(\Omega)} + \|\nabla(|\boldsymbol{e}_u| |\boldsymbol{\xi}_u|)\|_{L^1(\Omega)}) \\ &\leq c_S \|\boldsymbol{u}\|_{L^\infty(S_1)} (1 + \nu^{-1})(\|\boldsymbol{\xi}_u\|^2 + \|\boldsymbol{\eta}_u\|^2) + \epsilon(|||\boldsymbol{\xi}_u|||^2_{\boldsymbol{u}_h} + |||\boldsymbol{\eta}_u|||^2_{\boldsymbol{u}_h}). \end{aligned}$$

Hence, the sum of the two terms T_1 and T_2 can now be bounded by

$$T \leq C_G(\|\boldsymbol{\xi}_u\|^2 + \|\boldsymbol{\eta}_u\|^2) + (\epsilon + c_\beta \epsilon^{-1})|||\boldsymbol{\eta}_u|||^2_{\boldsymbol{u}_h} + \epsilon |||\boldsymbol{\xi}_u|||^2_{\boldsymbol{u}_h},$$

with a still arbitrary parameter $\epsilon > 0$. In combination with the upper bound for R we arrive for $\epsilon = \frac{1}{4}$ at

$$\partial_t \|\boldsymbol{\xi}_u\|^2 + |||\boldsymbol{\xi}_u|||_{\boldsymbol{u}_h}^2 \leq \|\partial_t \boldsymbol{\eta}_u\|^2 + 4 \sum_M \phi_M + C_G(\|\boldsymbol{\xi}_u\|^2 + \|\boldsymbol{\eta}_u\|^2) + c'_\beta |||\boldsymbol{\eta}_u|||_{\boldsymbol{u}_h}^2.$$

Application of the Gronwall lemma yields the assertion. □

Remark 3 If the solution $\boldsymbol{u}$ is sufficiently smooth, the previous Theorem can be used to derive an upper bound of the projection error in terms of powers of h_M by using

$$\phi_M(\tau) \leq C h_M^{2k} \left(\|\partial_t \boldsymbol{u}(\tau)\|_{H^2(M)}^2 + (c_5 + c_6)\|\boldsymbol{u}(\tau)\|_{H^{k+1}(M)}^2 + c_4 \|p(\tau)\|_{H^k(M)}^2 \right)$$

with $c_4 = (\nu + \gamma_M)^{-1}$, $c_5 = 1 + \nu + \gamma_{max} + \alpha_0$, and $c_6 = h_M^2 \nu^{-1} \|\boldsymbol{u}\|_{L^\infty(M)}^2$.

Remark 4 The term c_6 in the previous Remark can be avoided by using a different bound in the proof of Theorem 2, see Dallmann [7] (page 44). However, this leads to a larger Gronwall constant $\widetilde{C}_G = C_G + \|\boldsymbol{u}_h\|_{L^\infty(\Omega)}^2$.

5 Numerical Results

We want to support the above analysis by numerical examples that show the desired convergence results in space. In particular we like to see that the error does not blow up in time or space, even if there is inflow at the boundary S_1.

The considered domain is given by $\Omega := (0, 2\pi) \times (-\pi, \pi)$. We use the directional do-nothing (DDN) at $S_1 := \{(2\pi, y) : -\pi \leq y \leq \pi\}$ with the parameter $\beta = 1$, and Dirichlet boundary at $S_0 := \partial\Omega \setminus S_1$. Let $\chi : \mathbb{R} \to [0, 1]$ defined as $\chi(y) = 1$ if $y < 0$, and $\chi(y) = 0$ for $y \geq 0$. The exact solution in analytical form and the corresponding right hand side are given by

$$\begin{aligned}
\boldsymbol{u}(x, y) &= (\sin(y)\cos(t)^2, 0)^T, \\
p(x, y) &= -\tfrac{1}{2}\chi(y)\sin(y)^2\cos(t)^4, \\
\boldsymbol{f}(x, y) &= (-\sin(2t)\sin(y) + \cos(t)^2\sin(y)\nu, -\chi(y)\cos(y)\sin(y)\cos(t)^4)^T.
\end{aligned}$$

We investigate the convergence behavior for the classical Taylor-Hood pair Q_2/Q_1. Since we are not interested in the error due to time discretization we set $\Delta t = 10^{-4}$ and evaluate the error at $T = 10^{-2}$.

In Fig. 1 we depict the L^2-errors with respect to velocity and pressure, $\|\boldsymbol{u} - \boldsymbol{u}_h\|$ and $\|p - p_h\|$, in dependence of a uniform mesh size h for various viscosities ν. We compare with ($\gamma = 1$) and without ($\gamma = 0$) div-div stabilization. For the velocity error in L^2 we observe convergence of third order in the case $\gamma = 1$. Without div-div stabilization the convergence order of $\|\boldsymbol{u} - \boldsymbol{u}_h\|$ is reduced. For the pressure, second

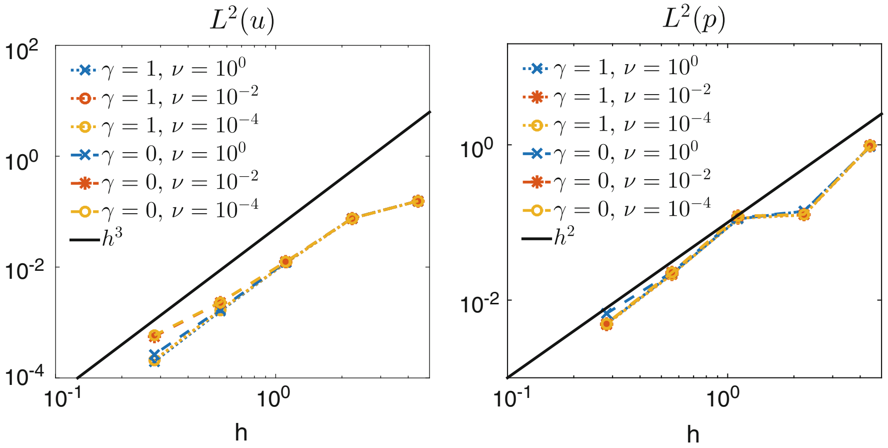


Fig. 1 L^2-errors of $\boldsymbol{u}$ and p for Taylor-Hood ($\mathscr{Q}_2/\mathscr{Q}_1$) elements

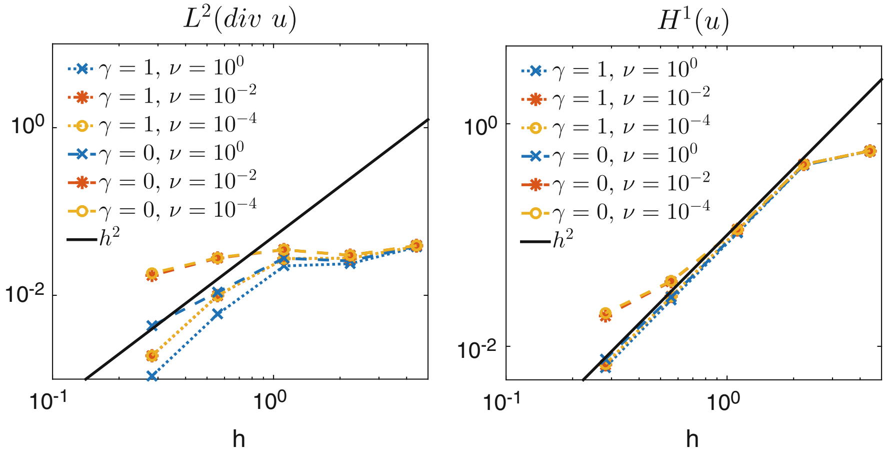


Fig. 2 Errors $\|\text{div}\,(\boldsymbol{u} - \boldsymbol{u}_h)\|$ and $|\boldsymbol{u} - \boldsymbol{u}_h|_1$ for Taylor-Hood ($\mathscr{Q}_2/\mathscr{Q}_1$) elements

order convergence can be observed which is in line with our analysis. The pressure error does essentially not depend on any of the parameters.

In Fig. 2 we show the errors $|\boldsymbol{u} - \boldsymbol{u}_h|_1$, and $\|\text{div}\,(\boldsymbol{u} - \boldsymbol{u}_h)\|$. Both quantities show quadratic convergence, i.e. at optimal rate, if div-div stabilization is used. For the velocity energy error $\|\boldsymbol{u} - \boldsymbol{u}_h\|$ and the $H^1(\Omega)$ error the results deviate from the optimal rate of convergence (h^3 resp. h^2) if no div-div stabilization is used. However, the biggest impact of the stabilization can be seen for the divergence error $\|\text{div}\,(\boldsymbol{u} - \boldsymbol{u}_h)\|$. For sufficiently small viscosity the error stays nearly constant if no div-div stabilization is used. Optimal convergence rates can be recovered if div-div stabilization is used. With respect to the LPS stabilization we did not observe any significant influence for the considered norms. Compared to results in [1] for Dirichlet boundary conditions the div-div stabilization seems to play a more important role.

References

1. D. Arndt, H. Dallmann, G. Lube, Local projection FEM stabilization for the time-dependent incompressible Navier–Stokes problem. Numer. Methods Partial Differ. Equ. **31**(4), 1224–1250 (2015)
2. Y. Bazilevs, J. Gohean, T. Hughes, R. Moser, Y. Zhang, Patient-specific isogeometric fluid–structure interaction analysis of thoracic aortic blood flow due to implantation of the Jarvik 2000 left ventricular assist device. Comput. Methods Appl. Mech. Eng. **198**(45), 3534–3550 (2009)
3. M. Braack, Outflow condition for the Navier–Stokes equations with skew-symmetric formulation of the convective term, in *BAIL 2014*. Lecture Notes in Computational Science and Engineering (Springer, Cham, 2016), pp. 35–45
4. M. Braack, P.B. Mucha, Directional do-nothing condition for the Navier–Stokes equations. J. Comput. Math. **32**(5), 507–521 (2014)
5. C.-H. Bruneau, P. Fabrie, New efficient boundary conditions for incompressible Navier–Stokes equations: a well-posedness result. RAIRO-Modélisation mathématique et analyse numérique **30**(7), 815–840 (1996)
6. C. Bruneau, P. Fabrie et al., Effective downstream boundary conditions for incompressible Navier–Stokes equations. Int. J. Numer. Methods Fluids **19**(8), 693–705 (1994)
7. H. Dallmann, Finie element methods with local projection stabilization for thermally coupled incompressible flow, Ph.D. thesis, University of Göttingen, 2015
8. S. Dong, A convective-like energy-stable open boundary condition for simulations of incompressible flows. J. Comput. Phys. **302**, 300–328 (2015)
9. V. Girault, L.R. Scott, A quasi-local interpolation operator preserving the discrete divergence. Calcolo **40**(1), 1–19 (2003)
10. J. Heywood, R. Rannacher, S. Turek, Artificial boundaries and flux and pressure conditions for the incompressible Navier–Stokes equations. Int. J. Numer. Methods Fluids **22**, 325–352 (1996)
11. G. Matthies, L. Tobiska, Local projection type stabilization applied to inf–sup stable discretizations of the Oseen problem. IMA J. Numer. Anal. **35**(1), 239–269 (2015)
12. M.E. Moghadam, Y. Bazilevs, T.-Y. Hsia, I.E. Vignon-Clementel, A.L. Marsden et al., A comparison of outlet boundary treatments for prevention of backflow divergence with relevance to blood flow simulations. Comput. Mech. **48**(3), 277–291 (2011)

Quasi-Optimality Constants for Parabolic Galerkin Approximation in Space

Francesca Tantardini and Andreas Veeser

Abstract We consider Galerkin approximation in space of linear parabolic initial-boundary value problems where the elliptic operator is symmetric and thus induces an energy norm. For two related variational settings, we show that the quasi-optimality constant equals the stability constant of the L^2-projection with respect to that energy norm.

1 Introduction

A Galerkin method S for a variational problem is quasi-optimal in a norm $\|\cdot\|$ if there exists a constant q such that

$$\|u - U_S\| \leq q \inf_v \|u - v\|, \tag{1}$$

where u is any variational solution, U_S its associated Galerkin approximation and v varies in the discrete trial space. The quasi-optimality constant q_S is the best constant q in (1), and thus measures how well the Galerkin method S exploits the approximation potential offered by the discrete trial space. The determination or estimation of q_S is therefore the ideal first step in an a priori error analysis.

F. Tantardini (✉)
Fakultät für Mathematik, Ruhr-Universität Bochum, Universitätsstraße 150, 44801 Bochum, Germany
e-mail: francesca.tantardini@rub.de

A. Veeser
Dipartimento di Matematica, Università degli Studi di Milano, via Saldini 50, 20133 Milano, Italy
e-mail: andreas.veeser@unimi.it

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_11

Here we are interested in Galerkin approximation in space for linear parabolic initial-boundary value problems like

$$\begin{aligned} \partial_t u - \Delta u = f \text{ in } \Omega \times (0,T), \\ u = 0 \text{ on } \partial\Omega \times (0,T), \quad u(\cdot,0) = w \text{ in } \Omega. \end{aligned} \tag{2}$$

Whereas for the stationary case, i.e. elliptic problems, quasi-optimality results like Céa's lemma are very common, such results have been less explored for parabolic problems. A common assumption of such results is that the L^2-projection onto the underlying discrete space is H^1-stable; see, e.g., [4, 5, 7], where the norm in (1) is either the one of $H^1(H^{-1}) \cap L^2(H^1)$ or the one of $L^2(H^1)$. Recently, the authors [8] have clarified the role of this assumption by showing that it is also necessary. This follows by applying the inf-sup theory [2, 3] to two weak, essentially dual formulations: the standard weak formulation with trial space $H^1(H^{-1}) \cap L^2(H^1)$ and the ultra-weak formulation with trial space $L^2(H^1)$.

This short note underlines the close relationship between parabolic quasi-optimality and the H^1-stability of the L^2-projection. It improves the results of [8] in the special case of a time-independent symmetric elliptic operator. For the model problem (2) and both variational formulations, this improvement reads as follows: the quasi-optimality constant of a Galerkin approximation with values in a discrete subspace S of H_0^1 is given by the operator norm in H_0^1 of the L^2-projection onto S:

$$q_{\mathrm{std};S} = \|P_S\|_{\mathscr{L}(H_0^1)} = q_{\mathrm{ult};S}. \tag{3}$$

2 Petrov-Galerkin Framework and Quasi-Optimality

This section, which is taken from [8], provides the general framework for the derivation of our quasi-optimality results. Let $(H_1, \|\cdot\|_1)$ and $(H_2, \|\cdot\|_2)$ be two real Hilbert spaces. The dual space H_2^* of H_2 is equipped with the usual dual norm $\|\ell\|_{H_2^*} = \sup_{\|\varphi\|_2=1} \ell(\varphi)$ for $\ell \in H_2^*$. Moreover, let b be a real-valued bounded bilinear form on $H_1 \times H_2$ and set $C_b := \sup_{\|v\|_1=\|\varphi\|_2=1} |b(v,\varphi)|$. We consider the problem

$$\text{given } \ell \in H_2^*, \text{ find } u \in H_1 \text{ such that } \forall \varphi \in H_2 \quad b(u,\varphi) = \ell(\varphi) \tag{4}$$

and say that it is well-posed if, for any $\ell \in H_2^*$, there exists a unique solution that continuously depends on ℓ. This holds if and only if there hold the following two conditions involving the so-called inf-sup constant c_b, cf. [3]:

$$c_b := \inf_{\|v\|_1=1} \sup_{\|\varphi\|_2=1} b(v,\varphi) > 0 \qquad \text{(uniqueness)}, \tag{5a}$$

$$\forall \varphi \in H_2 \setminus \{0\} \, \exists v \in H_1 \quad b(v,\varphi) > 0 \qquad \text{(existence)}. \tag{5b}$$

If (5) is satisfied, we have the duality

$$\inf_{\|v\|_1=1} \sup_{\|\varphi\|_2=1} b(v,\varphi) = \inf_{\|\varphi\|_2=1} \sup_{\|v\|_1=1} b(v,\varphi). \tag{6}$$

For notational simplicity, we take the viewpoint that a Petrov-Galerkin method for problem (4) is characterized by one pair of subspaces, instead of a family of pairs. Let $M_i \subset H_i$, $i = 1, 2$, be nontrivial and proper subspaces. The Petrov-Galerkin method $M = (M_1, M_2)$ for (4) reads

$$\text{given } \ell \in H_2^*, \text{ find } U_M \in M_1 \text{ such that } \forall \varphi \in M_2 \quad b(U_M, \varphi) = \ell(\varphi). \tag{7}$$

Problem (7) is well-posed if and only if there hold the semidiscrete counterparts of (5), involving the semidiscrete inf-sup constant c_M:

$$c_M := \inf_{v\in M_1:\|v\|_1=1} \sup_{\varphi\in M_2:\|\varphi\|_2=1} b(v,\varphi) > 0,$$

$$\forall \varphi \in M_2 \setminus \{0\}\ \exists v \in M_1 \quad b(v,\varphi) > 0.$$

A method M is quasi-optimal if there exists a constant $q \geq 1$ such that, for any $\ell \in H_2^*$, there holds

$$\|u - U_M\|_1 \leq q \inf_{v\in M_1} \|u - v\|_1 . \tag{8}$$

The quasi-optimality constant q_M of the method M is the smallest constant verifying (8). The formula for q_M in [8, Theorem 2.1] or combining [2, 3] with [9] imply

$$\frac{c_b}{c_M} \leq q_M \leq \frac{C_b}{c_M}. \tag{9}$$

3 Two Weak Formulations of Linear Parabolic Problems

In order to cast parabolic initial-boundary value problems in the form (4), we briefly recall two suitable weak formulations thereof.

Let V and W be two separable Hilbert spaces such that $V \subset W \subset V^*$ forms a Hilbert triplet. The scalar product in W as well as the duality pairing of $V^* \times V$ is denoted by $\langle\cdot,\cdot\rangle$. The norms are indicated by $\|\cdot\|_V$, $\|\cdot\|_W$, and $\|\cdot\|_{V^*} = \sup_{\|v\|_V=1} \langle\cdot, v\rangle$.

Let $A \in \mathscr{L}(V, V^*)$ be a linear and continuous operator arising from a symmetric bilinear form a via $\langle Av, \varphi\rangle = a(v, \varphi)$. We assume that a is bounded and coercive, i.e.

$$\nu_a := \inf_{\|v\|_V=1} a(v, v) > 0, \qquad C_a := \sup_{\|v\|_V=\|\varphi\|_V=1} a(v, \varphi) < \infty. \tag{10}$$

In view of (10) and the symmetry of a, the energy norm $\|\cdot\|_a = \langle A\cdot, \cdot\rangle^{1/2}$ and the dual energy norm $\|\cdot\|_{a;*} := \sup_{\|\varphi\|_a=1} \langle \cdot, \varphi\rangle$ are equivalent to $\|\cdot\|_V$ and $\|\cdot\|_{V^*}$, respectively. Moreover, for every $\ell \in V^*$ we have

$$\|\ell\|_{a;*} = \sup_{\|\varphi\|_a=1} \left\langle A\varphi, A^{-1}\ell\right\rangle = \left\|A^{-1}\ell\right\|_a = \sqrt{\langle \ell, A^{-1}\ell\rangle}. \tag{11}$$

Finally, given a final time $T > 0$ and a Hilbert space X, we set $I := (0, T)$ and denote with $L^2(X) := L^2(I; X)$ the space of all Lebesgue-measurable and square-integrable functions of the form $I \to X$. In addition, if Y is another Hilbert space, we set $H^1(X, Y) := \{v \in L^2(X) \mid v' \in L^2(Y)\}$ and write $H^1(X)$ for $H^1(X, X)$.

3.1 Standard Weak Formulation

The standard weak formulation is very common, also for some nonlinear parabolic problems. In the above setting, it reads

$$\begin{gathered}\text{given} f \in L^2(V^*) \text{ and } w \in W, \text{ find } u \in H^1(V, V^*) \text{ such that} \\ u' + Au = f \text{ in } I, \quad u(0) = w\end{gathered} \tag{12}$$

and can be cast in the form (4) by choosing $H_1 = H^1(V, V^*)$ and $H_2 = \{\varphi = (\varphi_0, \varphi_1) \mid \varphi_0 \in W, \varphi_1 \in L^2(V)\}$ with norms

$$\|v\|_1^2 = \|v(T)\|_W^2 + \int_I \|v\|_a^2 + \left\|v'\right\|_{a;*}^2, \quad \|\varphi\|_2^2 = \|\varphi_0\|_W^2 + \int_I \|\varphi_1\|_a^2. \tag{13}$$

Bilinear form and right-hand side are given, respectively, by

$$b(v, \varphi) = b_{\mathrm{std}}(A; v, \varphi) := \langle v(0), \varphi_0\rangle + \int_I \left\langle v', \varphi_1\right\rangle + \langle Av, \varphi_1\rangle \tag{14}$$

and $\ell(\varphi) = \langle w, \varphi_0\rangle + \int_I \langle f, \varphi_1\rangle$. We denote the constants of b_{std} by C_{std} etc.

The norm $\|\cdot\|_1$ in (13) slightly differs from the corresponding definition in [8] because it involves $v(T)$ instead of $v(0)$. This modification offers the following advantage, which was already observed in [1]: the norms in (13) mimic the energy norm for a linear elliptic problem in that the operator $v \mapsto b(v, \cdot)$ is an isometry. We provide a proof because its arguments will be used in what follows.

Proposition 1 (Isometry) *For every $v \in H_1$, we have $\|b(v, \cdot)\|_{H_2^*} = \|v\|_1$.*

Proof In view of $\int_I \langle v', v \rangle = \|v(T)\|_W^2 - \|v(0)\|_W^2$, the symmetry of A and (11), we have the identity

$$\|v(0)\|_W^2 + \int_I \left\|A^{-1}v' + v\right\|_a^2 = \|v(T)\|_W^2 + \int_I \left\|A^{-1}v'\right\|_a^2 + \|v\|_a^2 = \|v\|_1^2 \tag{15}$$

for every $v \in H_1$. On the one hand, this gives, for every $v \in H_1$, $\varphi \in H_2$,

$$\begin{aligned} b(v, \varphi) &= \langle v(0), \varphi_0 \rangle + \int_I \left\langle v', \varphi_1 \right\rangle + \langle Av, \varphi_1 \rangle \\ &\le \left(\|v(0)\|_W^2 + \int_I \left\|A^{-1}v' + v\right\|_a^2 \right)^{1/2} \|\varphi\|_2 = \|v\|_1 \|\varphi\|_2 , \end{aligned}$$

which implies $\|b(v, \cdot)\|_{H_2^*} \le \|v\|_1$. On the other hand, choosing

$$\varphi_0 = v(0), \qquad \varphi_1 = v' + A^{-1}v \tag{16}$$

and using again (15), we get $\|\varphi\|_2 = \|v\|_1$ and

$$b(v, \varphi) = \|v(0)\|_W^2 + \int_I \left\langle v', v \right\rangle + \left\langle v', A^{-1}v \right\rangle + \langle Av, v \rangle = \|v\|_1^2 .$$

Hence, $\|b(v, \cdot)\|_{H_2^*} \ge \|v\|_1$. □

Corollary 2 (Standard bilinear form) *The bilinear form b in (14) is continuous and satisfies the inf-sup condition with $C_{\text{std}} = c_{\text{std}} = 1$.*

Proof The equalities follow readily from Proposition 1. The proof of the non-degeneracy condition (5b) can be found in [8, Prop. 3.1]. □

3.2 Ultra-Weak Formulation

Discontinuous Galerkin methods, applications in optimization and stochastic PDEs motivate to consider solution notions with less regularity in time. In order to obtain such a solution notion for (12), one may multiply the differential equation with a test function

$$\varphi \in H_T^1(V, V^*) := \{\varphi \in L^2(I; V) \mid \varphi' \in L^2(I, V^*), \varphi(T) = 0\},$$

integrate in time and by parts. This results in the ultra-weak formulation, which can be cast in the form (4) by choosing $H_1 = L^2(V)$, and $H_2 = H^1_T(V, V^*)$, with norms

$$\|v\|_1^2 = \int_I \|v\|_V^2, \qquad \|\varphi\|_2^2 = \int_I \|\varphi\|_a^2 + \|\varphi'\|_{a;*}^2.$$

Here, bilinear form and right-hand side are given, respectively, by

$$b(v, \varphi) = b_{\mathrm{ult}}(A; v, \varphi) := \int_I -\langle \varphi', v\rangle + \langle Av, \varphi\rangle \tag{17}$$

and $\ell(\varphi) = \langle w, \varphi(0)\rangle + \int_I \langle f, \varphi\rangle + \langle \varphi', f_1\rangle$, with $f \in L^2(V^*)$, $f_1 \in L^2(V)$ and $w \in W$. We denote the constants of b_{ult} by C_{ult} etc. Every solution of the standard weak formulation is one of the ultra-weak formulation.

Corollary 3 (Ultra-weak bilinear form) *The bilinear form b in (17) is continuous and satisfies the inf-sup condition with $C_{\mathrm{ult}} = c_{\mathrm{ult}} = 1$.*

Proof We exploit the duality with the standard weak formulation. Setting $\iota v(t) := v(T - t)$, $t \in I = (0, T)$ and using the symmetry of A, we have

$$\forall v_1 \in L^2(V), v_2 \in H^1_T(V, V^*) \quad b_{\mathrm{ult}}(A; v_1, v_2) = b_{\mathrm{std}}(A; \iota v_2, \iota v_1); \tag{18}$$

see [8, Lemma 4.1]. Since Proposition 1 holds also with $H^1_0(V, V^*) := \{v \in H^1(V, V^*) \mid v(0) = 0\}$ in place of $H^1(V, V^*)$, we thus deduce $C_{\mathrm{ult}} = C_{\mathrm{std}} = 1$ and $c_{\mathrm{ult}} = c_{\mathrm{std}} = 1$ with the help of (6). □

4 Galerkin Approximation in Space and Quasi-Optimality Constants

We review Galerkin approximation in space for the standard and the ultra-weak formulation and then derive identities for the corresponding quasi-optimality constants.

Let S be a finite-dimensional, nontrivial, and proper subspace of V. Observe that S is also a subspace of W and, with the identification $S^* = S$, also of V^*. As a subspace of V^*, we can equip $S = S^*$ with

$$\|\ell\|_{a;*} = \sup_{\varphi \in V} \frac{\langle \ell, \varphi\rangle}{\|\varphi\|_a} \quad \text{as well as} \quad \|\ell\|_{a;S^*} := \sup_{\varphi \in S} \frac{\langle \ell, \varphi\rangle}{\|\varphi\|_a}.$$

The following relationship, which can be found, e.g., in [8, Proposition 2.5], will be crucial:

$$\sup_{\ell \in S} \frac{\|\ell\|_{a;*}}{\|\ell\|_{a;S^*}} = \|P_S\|_{\mathscr{L}(V, \|\cdot\|_a)} := \sup_{\|w\|_a = 1} \|P_S w\|_a, \tag{19}$$

where P_S is the W-orthogonal projection onto S satisfying $\langle P_S w, \varphi\rangle = \langle w, \varphi\rangle$ for all $\varphi \in S$ and every $w \in W$.

4.1 Standard Weak Formulation

We first consider the standard weak formulation and define the spaces H_1, H_2, their norms and the bilinear form b as in Sect. 3.1. The Galerkin approximation with values in S is characterized by (7) with $M = (M_1, M_2)$ where

$$M_1 = H^1(S) \subset H_1, \qquad M_2 = S \times L^2(S) \subset H_2. \tag{20}$$

In order to determine the associated inf-sup constant $c_{\mathrm{std};S}$ in (5a), we first derive a discrete counterpart of Proposition 1. To this end, we define on M_1 the following S-dependent variant of $\|\cdot\|_1$:

$$\|v\|_{1;S}^2 := \|v(T)\|_W^2 + \int_I \|v\|_a^2 + \left\|v'\right\|_{a;S^*}^2 ,$$

where we replaced the dual norm $\|\cdot\|_{a;*}$ of the time derivative with the discrete dual norm $\|\cdot\|_{a;S^*}$. This gives rise to

$$\tilde{c}_{\mathrm{std};S} := \inf_{v\in M_1} \sup_{\varphi\in M_2} \frac{b(v,\varphi)}{\|v\|_{1;S}\,\|\varphi\|_2}, \quad \tilde{C}_{\mathrm{std};S} := \sup_{v\in M_1} \sup_{\varphi\in M_2} \frac{b(v,\varphi)}{\|v\|_{1;S}\,\|\varphi\|_2}$$

and

$$\inf_{v\in M_1} \frac{\|v\|_{1;S}}{\|v\|_1}\, \tilde{c}_{\mathrm{std};S} \le c_{\mathrm{std};S} \le \inf_{v\in M_1} \frac{\|v\|_{1;S}}{\|v\|_1}\, \tilde{C}_{\mathrm{std};S}. \tag{21}$$

Proposition 4 (Discrete isometry) *For every $v \in M_1$, we have*

$$\|b(v,\cdot)\|_{M_2^*} := \sup_{\varphi\in M_2} \frac{b(v,\varphi)}{\|\varphi\|_2} = \|v\|_{1;S}\,.$$

Proof In order to proceed as in the proof of Proposition 1, we introduce the discrete counterpart of A, namely the operator $A_S : S \to S^*$ given by $\langle A_S v, \varphi\rangle = a(v,\varphi)$, for every $v, \varphi \in S$. In analogy to (11), we have $\left\langle \ell, A_S^{-1}\ell\right\rangle = \|A_S^{-1}\ell\|_a^2 = \|\ell\|_{a;S^*}^2$. We thus conclude as in the proof of Proposition 1, upon replacing $\varphi = (\varphi_0, \varphi_1)$ in (16) with $\varphi_0 = v(0) \in S$, $\varphi_1 = v + A_S^{-1}v' \in L^2(S)$. □

Consequently, the counterparts of the identities in Corollary 2 are

$$\tilde{c}_{\mathrm{std};S} = \tilde{C}_{\mathrm{std};S} = 1, \tag{22}$$

which imply a symmetric error estimate for $\|\cdot\|_{1;S}$, similar to the one in [6]. For $\|\cdot\|_1$ instead, we have:

Theorem 5 (Quasi-optimality in $H^1(V, \|\cdot\|_a; V^*, \|\cdot\|_{a;*})$**)** *The quasi-optimality constant of the Galerkin method* (20) *is given in terms of the W-projection onto S by*

$$q_{\mathrm{std};S} = \|P_S\|_{\mathscr{L}(V,\|\cdot\|_a)} .$$

Proof Identity (19) entails that the ratio of the two norms in the trial space is

$$\sup_{v \in M_1} \frac{\|v\|_1}{\|v\|_{1;S}} = \|P_S\|_{\mathscr{L}(V,\|\cdot\|_a)} , \tag{23}$$

see [8, Proposition 2.5 and (3.14)]. We thus deduce

$$q_{\mathrm{std};S} = c_{\mathrm{std};S}^{-1} = \sup_{v \in M_1} \frac{\|v\|_1}{\|v\|_{1;S}} = \|P_S\|_{\mathscr{L}(V,\|\cdot\|_a)} . \tag{24}$$

by using Corollary 2 in (9) and (22) in (21). □

Remark 6 (Non-symmetric case) If a is not symmetric, Theorem 5 can be generalized to

$$\kappa_a^{-1} \|P_S\|_{\mathscr{L}(V,\|\cdot\|_a)} \leq q_{\mathrm{std};S} \leq \kappa_a \|P_S\|_{\mathscr{L}(V,\|\cdot\|_a)} ,$$

where $\|\cdot\|_a$ is given by the symmetric part of a and κ_a depends on C_a and ν_a, with $\kappa_a = 1$ whenever a is symmetric. To this end, the bilinear form is split into its symmetric and skew-symmetric part, where the latter part is treated as a perturbation. An alternative and more general approach is offered by [8]. That analysis appears to be simpler but we only have $\kappa_a = \sqrt{2}$ if a is symmetric and one adopts the above energy-norm setting.

4.2 Ultra-Weak Formulation

We turn to Galerkin approximation based upon the ultra-weak formulation. Let the spaces H_1, H_2, their norms and the bilinear form b be given as in Sect. 3.2. The corresponding Galerkin approximation with values in S is characterized by (7) with $M = (M_1, M_2)$ where

$$M_1 = L^2(S) \subset H_1, \quad M_2 = H_T^1(S) := H^1(S) \cap H_T^1(V, V^*) \subset H_2. \tag{25}$$

Also, the Galerkin approximation of the ultra-weak formulation generalizes the Galerkin approximation of the standard weak formulation. Moreover:

Theorem 7 (Quasi-optimality in $L^2(V, \|\cdot\|_a)$) *The quasi-optimality constant of the ultra-weak Galerkin method* (25) *is determined in terms of the W-projection onto S by*

$$q_{\mathrm{ult};S} = \|P_S\|_{\mathscr{L}(V,\|\cdot\|_a)} \,.$$

Proof We exploit again duality. To this end, notice first that Proposition 4 and (23) hold also if $H^1(S)$ is replaced by $H^1_0(S) := \{v \in H^1(S) \mid v(0) = 0\}$. Hence, the discrete inf-sup constant does not change under this replacement and (18) yields $c_{\mathrm{ult};S} = c_{\mathrm{std};S}$. We thus obtain

$$q_{\mathrm{ult};S} = c^{-1}_{\mathrm{ult};S} = c^{-1}_{\mathrm{std};S} = \|P_S\|_{\mathscr{L}(V,\|\cdot\|_a)}$$

by using Corollary 3 in (9) and (24). □

Theorems 5 and 7 with $W = L^2(\Omega)$, $V = H^1_0(\Omega)$ and $A = -\Delta$ yield (3).

References

1. R. Andreev, Wavelet-in-time multigrid-in-space preconditioning of parabolic evolution equations. SIAM J. Sci. Comput. **38**(1), A216-A242 (2016)
2. D.N. Arnold, I. Babuška, J. Osborn, Finite element methods: principles for their selection. Comput. Methods Appl. Mech. Eng. **45**(1–3), 57–96 (1984)
3. I. Babuška, Error-bounds for finite element method. Numer. Math. **16**, 322–333 (1970/1971)
4. I. Babuška, T. Janik, The *h-p* version of the finite element method for parabolic equations. I. The *p*-version in time. Numer. Methods Partial Differ. Equ. **5**(4), 363–399 (1989)
5. K. Chrysafinos, L.S. Hou, Error estimates for semidiscrete finite element approximations of linear and semilinear parabolic equations under minimal regularity assumptions. SIAM J. Numer. Anal. **40**(1), 282–306 (2002)
6. T. Dupont, Mesh modification for evolution equations. Math. Comput. **39**(159), 85–107 (1982)
7. W. Hackbusch, Optimal $H^{p,p/2}$ error estimates for a parabolic Galerkin method. SIAM J. Numer. Anal. **18**(4), 681–692 (1981)
8. F. Tantardini, A. Veeser, The L^2-projection and quasi-optimality in Galerkin methods for parabolic equations. SIAM J. Numer. Anal. **54** (1), 317-340 (2016)
9. J. Xu, L. Zikatanov, Some observations on Babuška and Brezzi theories. Numer. Math. **94**(1), 195–202 (2003)

Numerical Studies on a Second Order Explicitly Decoupled Variational Multiscale Method

Mine Akbas, Songul Kaya, and Leo Rebholz

Abstract Projection based variational multiscale (VMS) methods are a very successful technique in the numerical simulation of high Reynolds number flow problems using coarse discretizations. However, their implementation into an existing (legacy) codes can be very challenging in practice. We propose a second order variant of projection-based VMS method for non-isothermal flow problems. The method adds stabilization as a decoupled post-processing step for both velocity and temperature, and thus can be efficiently and easily used with existing codes. In this work, we propose the algorithm and give numerical results for convergence rates tests and coarse mesh simulation of Marsigli flow.

1 Introduction

We consider the Boussinesq system on an open, simply connected domain $\Omega \subset \mathbb{R}^d$, $d = 2$ or 3, with boundary $\partial\Omega$ subject to no slip boundary conditions

$$\mathbf{u}_t - \nu\Delta\mathbf{u} + (\mathbf{u}\cdot\nabla)\mathbf{u} + \nabla p = Ri\langle\mathbf{0},\theta\rangle + \mathbf{f}, \tag{1}$$

$$\nabla\cdot\mathbf{u} = 0, \quad \text{in } \Omega \tag{2}$$

$$\theta_t - \kappa\Delta\theta + (\mathbf{u}\cdot\nabla)\theta = \gamma, \tag{3}$$

$$\mathbf{u}|_{t=0} = \mathbf{0} \qquad \theta|_{t=0} = 0, \quad \text{on } \partial\Omega, \tag{4}$$

where the Richardson number is denoted by Ri, $Ri\langle\mathbf{0},\theta\rangle$ is a vector, $\mathbf{u}$ is the fluid velocity, p the pressure, θ the temperature, and $\mathbf{f}$ and γ are the prescribed forcing

M. Akbas • S. Kaya (✉)
Department of Mathematics, Middle East Technical University, 06800 Ankara, Turkey
e-mail: miakbasb@metu.edu.tr; smerdan@metu.edu.tr

L. Rebholz
Department of Mathematical Sciences, Clemson University, Clemson, SC, 29634 USA
e-mail: rebholz@clemson.edu

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_12

and source, respectively. The kinematic viscosity ν is inversely proportional to the Reynolds number, *Re*, and $\kappa = Re^{-1}Pr^{-1}$, where *Pr* is the Prandtl number.

This report proposes and tests a second order projection-based variational multiscale (VMS) method for non-isothermal flow simulations. The success of VMS methods and different realizations have been documented in many works [3–8]. Our proposed method adds stabilizations with the L^2 orthogonal projections for both velocity and temperature such that they truncate the scales at the optimal place, i.e. at the spacial mesh-width. These additional projection and stabilization steps are post-processing steps and thus can be used to incorporate VMS into existing codes. The novel idea of explicitly decoupled VMS methods was originally proposed for the NSE by Layton et al. in [9], and was extended to a first order method for non-isothermal flows in [1]. The purpose of this work is to extend idea from [1] to a second order method.

The paper is organized as follows. In Sect. 2, we present a second-order projection based VMS method for (1), (2), (3) and (4). Sect. 3 presents numerical experiments for the proposed algorithm. Finally, Sect. 4 is devoted to the conclusions of the paper.

2 Numerical Scheme

We consider conforming finite element approximations of velocity, pressure, and temperature in subspaces of $\boldsymbol{X} := [\boldsymbol{H}_0^1(\Omega)]^d, Q := L_0^2(\Omega), W := H_0^1(\Omega)$. The coarse and fine mesh are denoted by Π_H and Π_h. We assume that the finite element spaces satisfy the inf-sup compatibility condition.

We denote skew-symmetric trilinear forms by

$$b^*(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) := \frac{1}{2}[((\boldsymbol{u} \cdot \nabla)\boldsymbol{u}, \boldsymbol{w}) - ((\boldsymbol{u} \cdot \nabla)\boldsymbol{w}, \boldsymbol{v})], \tag{5}$$

$$c^*(\boldsymbol{u}, \theta, \chi) := \frac{1}{2}[((\boldsymbol{u} \cdot \nabla)\theta, \chi) - ((\boldsymbol{u} \cdot \nabla)\chi, \theta)], \tag{6}$$

and the L^2 orthogonal projections are defined by P_μ^H (where $\mu = \mathbf{u}, \theta$) into $L_H \subset L := (L^2(\Omega))^{d \times d}$ on the coarse mesh Π_H.

Remark 1 Numerical studies have shown the choice of L_H is crucial in numerical experiments. In this report, we assume that L_H is selected for both velocity and temperature so that it has lower order than $\boldsymbol{X}_h, Q_h$ and W_h, on the same grid.

We study the following second order 2-step fully discrete version of (1), (2), (3) and (4).

Algorithm Let $\boldsymbol{f} \in L^\infty(0, T; \boldsymbol{H}^{-1}(\Omega)^d), \gamma \in L^\infty(0, T; H^{-1}(\Omega))$, and $\boldsymbol{u}_0 \in \boldsymbol{L}^2(\Omega)^d$ and $\theta_h^0 \in L^2(\Omega)$. Choose a finite end time T, and time step $\Delta t > 0$ such that $T = M\Delta t$ and $t^{n+1} = (n+1)\Delta t$, $n = 1, 2, \ldots, M-1$. Denote the fully discrete solutions by

$$\boldsymbol{u}_h^{n+1} := \boldsymbol{u}_h(t^{n+1}), \qquad p_h^{n+1} := p_h(t^{n+1}), \qquad \theta_h^{n+1} := \theta_h(t^{n+1}),$$

for all $n = 1, 2, \dots, M-1$. Define $\boldsymbol{u}_h^{-1} = \boldsymbol{u}_h^0$ and $\theta_h^{-1} = \theta_h^0$ to be the nodal interpolants of $\boldsymbol{u}_0$ and θ_0, respectively. Then for user-selected eddy viscosity parameters are denoted by $\alpha_1 = \alpha_1(\mathbf{x}, h)$, $\alpha_2 = \alpha_2(x, h)$, find $(\boldsymbol{u}_h^{n+1}, p_h^{n+1}, \theta_h^{n+1}) \in (\boldsymbol{X}_h, Q_h, W_h)$ via the following two steps:

Step 1: Find $(\boldsymbol{w}_h^{n+1}, p_h^{n+1}, \phi_h^{n+1}) \in (\boldsymbol{X}_h, Q_h, W_h)$ such that

$$\frac{1}{\Delta t}\left(\boldsymbol{w}_h^{n+1} - \boldsymbol{u}_h^n, \boldsymbol{v}_h\right) + \nu(\nabla \boldsymbol{w}_h^{n+1/2}, \nabla \boldsymbol{v}_h) + b^*\left(\frac{3}{2}\boldsymbol{u}_h^n - \frac{1}{2}\boldsymbol{u}_h^{n-1}, \boldsymbol{w}_h^{n+1/2}, \boldsymbol{v}_h\right)$$

$$-(p_h^{n+1}, \nabla \cdot \boldsymbol{v}_h) = Ri((0, \frac{3}{2}\theta_h^n - \frac{1}{2}\theta_h^{n-1}), \boldsymbol{v}_h) + (\boldsymbol{f}(t^{n+1/2}), \boldsymbol{v}_h) \tag{7}$$

$$(\nabla \cdot \boldsymbol{w}_h^{n+1}, q_h) = 0, \tag{8}$$

$$\frac{1}{\Delta t}\left(\phi_h^{n+1} - \theta_h^n, \chi_h\right) + \kappa(\nabla \phi_h^{n+1/2}, \nabla \chi_h) + c^*\left(\frac{\boldsymbol{u}_h^n + \boldsymbol{w}_h^{n+1}}{2}, \phi_h^{n+1/2}, \chi_h\right)$$

$$= (\gamma(t^{n+1/2}), \chi_h). \tag{9}$$

for all $(\boldsymbol{v}_h, q_h, \chi_h) \in (\boldsymbol{X}_h, Q_h, W_h)$.
Step 2: Find $(\boldsymbol{u}_h^{n+1}, \lambda_h^{n+1}, \theta_h^{n+1}) \in (\boldsymbol{X}_h, Q_h, W_h)$ satisfying

$$\frac{1}{\Delta t}\left(\boldsymbol{w}_h^{n+1} - \boldsymbol{u}_h^{n+1}, \boldsymbol{\varphi}_h\right) = (\lambda_h^{n+1}, \nabla \cdot \boldsymbol{\varphi}_h) + \alpha_1\left(\nabla \boldsymbol{u}_h^{n+1}, \nabla \boldsymbol{\varphi}_h\right)$$

$$-\alpha_1\left(P_u^H \nabla \boldsymbol{w}_h^{n+1}, \nabla \boldsymbol{\varphi}_h\right) \tag{10}$$

$$(\nabla \cdot \boldsymbol{u}_h^{n+1}, r_h) = 0, \tag{11}$$

$$\frac{1}{\Delta t}\left(\phi_h^{n+1} - \theta_h^{n+1}, \psi_h\right) = \alpha_2\left(\nabla \theta_h^{n+1}, \nabla \psi_h\right) - \alpha_2\left(P_\theta^H \nabla \phi_h^{n+1}, \nabla \psi_h\right) \tag{12}$$

for all $(\boldsymbol{\varphi}_h, r_h, \psi_h) \in (\boldsymbol{X}_h, Q_h, W_h)$.

Remark 2 In our experiments, we choose $\alpha_1 = O(h^2)$ and $\alpha_2 = O(h^2)$. This is the choice that was used in the first order method, and was found successful. In future work, we plan to perform an analysis of the proposed method to see, among other things, what choices of α_1 and α_2 can be made that will provide optimal accuracy.

3 Numerical Experiments

In this section, we present two numerical experiments that demonstrate the performance of the method for the following two test problems:

3.1 Numerical Experiment 1: Convergence Rate Verification

The proposed numerical scheme for the Boussinesq system is expected to be a second order in time due to Crank-Nicolson-type timestepping, and a second order in space due to the stabilization method. To verify the order of convergence of the Algorithm, we choose the prescribed solution as

$$\begin{aligned} \boldsymbol{u} &= \begin{pmatrix} \cos(\pi y) \\ \sin(\pi x) \end{pmatrix} (1+\exp(t)), \\ p &= \sin(\pi(x+y))(1+\exp(t)), \\ \theta &= \sin(\pi x) + y\exp(t). \end{aligned}$$

We have chosen the parameters

$$Re = 1.0, \qquad Pr = 1.0, \qquad Ri = 1.0, \qquad \kappa = 1.0.$$

in our computations.

The right hand side functions $\boldsymbol{f}$ and γ are chosen such that $(\boldsymbol{u}, p, T)$ fulfill the Boussinesq system. We present computations with the finite element spaces $(\boldsymbol{P}_2, P_1, P_2)$, for the velocity, pressure and temperature, respectively. Here $\boldsymbol{P}_2$, is the space of continuous piecewise quadratic functions and P_1 is the space continuous piecewise linears. These conforming pairs of finite element spaces for the velocity and pressure satisfy the inf-sup condition. In addition to these finite element spaces, for the coarser space, we have chosen $\boldsymbol{L}_H = \boldsymbol{P}_1$ and the parameters are chosen as $H = h$, $\alpha_1 = \alpha_2 = h^2$. We have carried out all computations with the end time $T = 0.05$ and decreasing value of Δt and h. The errors are calculated in the following norm:

$$|||\boldsymbol{v}|||_{L^2(0,T;H^1(\Omega))} := \left(\Delta t \sum_{n=0}^{M-1} \|\nabla \boldsymbol{v}^{n+1}\|^2 \right)^{1/2}.$$

We present results for $|||\boldsymbol{u} - \boldsymbol{u}_h|||_{L^2(0,T;H^1(\Omega))}$ and $|||\theta - \theta_h|||_{L^2(0,T;H^1(\Omega))}$. The second order convergence of the errors in both velocity and temperature can be clearly observed in Table 1.

Table 1 Velocity and temperature convergence rate results with $T = 0.05$

Δt	h	$\|\|\boldsymbol{u} - \boldsymbol{u}_h\|\|_{L^2(0,T;H^1(\Omega))}$	Rate	$\|\|\theta - \theta_h\|\|_{L^2(0,T;H^1(\Omega))}$	Rate
$T/1$	1/2	1.306e−1	—	4.646e−2	—
$T/2$	1/4	3.387e−2	1.947	1.127e−2	2.044
$T/4$	1/8	8.875e−3	1.932	2.729e−3	2.046
$T/8$	1/16	2.289e−3	1.955	6.833e−4	1.998
$T/16$	1/32	5.838e−4	1.971	1.715e−4	1.994
$T/32$	1/64	1.477e−4	1.982	4.297e−5	1.997

3.2 Numerical Experiment 2: Marsigli's Flow with Re = 2,000

In this numerical experiment, we simulate a physical situation described by Marsigli in 1681, which reveals that when two fluids with different densities come across, the fluids with lower density tends to move onto the top of the fluid with higher density. This situation was simulated by H. Johnston et al. in [10], which is the study of the Boussinesq system with the fourth order finite difference scheme.

We consider the domain to be a box with width 8 and height 1, and prescribe no-slip boundary conditions for the velocity on all sides ($\boldsymbol{u} = \boldsymbol{0}$ on $\partial\Omega$), and perfect insulation on all sides that ($\nabla\theta \cdot n = 0$ on the boundary). The initial velocity is taken to be at rest, and the initial temperature is discontinuous, with $\theta = 1.5$ on the left half of the box ($x \leq 4$) and $\theta = 1$ for the right half ($x > 4$). We choose flow parameters of $Re = 2{,}000$, $Ri = 4$, and $Pr = 1$.

For the direct numerical simulation (DNS) of the no-model, we refer the results in [11]. In this work, the velocity, pressure and temperature spaces are chosen as $(\boldsymbol{P}_2, P_1, P_2)$. Then the approximate solutions of the no-model (no-model solutions mean that Step 1 solutions of Algorithm) are calculated with the time step $\Delta t = 0.002$ on a triangular mesh with the mesh-width $h = 0.01$, which corresponds to a grid 800 × 100. The main goal in our numerical experiment is to compare the approximate solutions of Algorithm with the no-model solutions on a coarse mesh, and our expectation is to get more accurate approximate solutions for the proposed algorithm. To do this, we take the time step $\Delta t = 0.002$ and choose finite element spaces $(\boldsymbol{P}_2, P_1, P_2)$. Then, we compute the solutions of no-model on a coarse mesh, which provides 14,762 degrees of freedom (d.o.f.) for the velocity , 1891 d.o.f. for the pressure and 7381 d.o.f. for the temperature at $t = 4$ and $t = 8$.

All computations used *FreeFem++* [2], and the results at $t = 4$ and $t = 8$ are presented in Fig. 1 for no-model, and for the decoupled VMS model in Fig. 2. Significant oscillations can be seen in velocity and temperature solutions for the no-model solution. However, the VMS solution provides a good qualitative solution without oscillations, which demonstrates that the algorithm is much more successful on the coarse mesh.

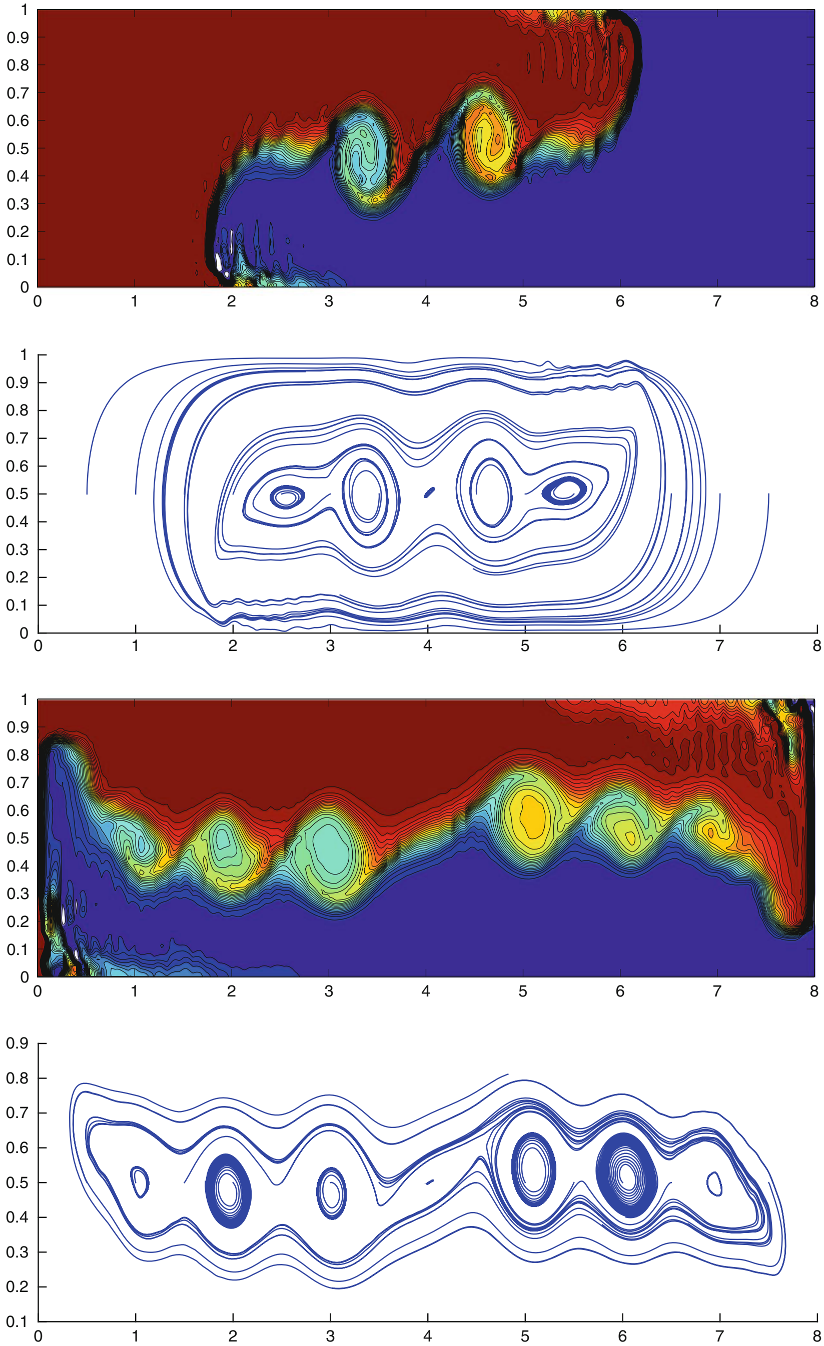


Fig. 1 Temperature contours and velocity streamlines for the $Re = 2{,}000$, $Ri = 4$, $Pr = 1$ Marsigli flow test with $t = 4$, and 8, respectively, with only Step 1

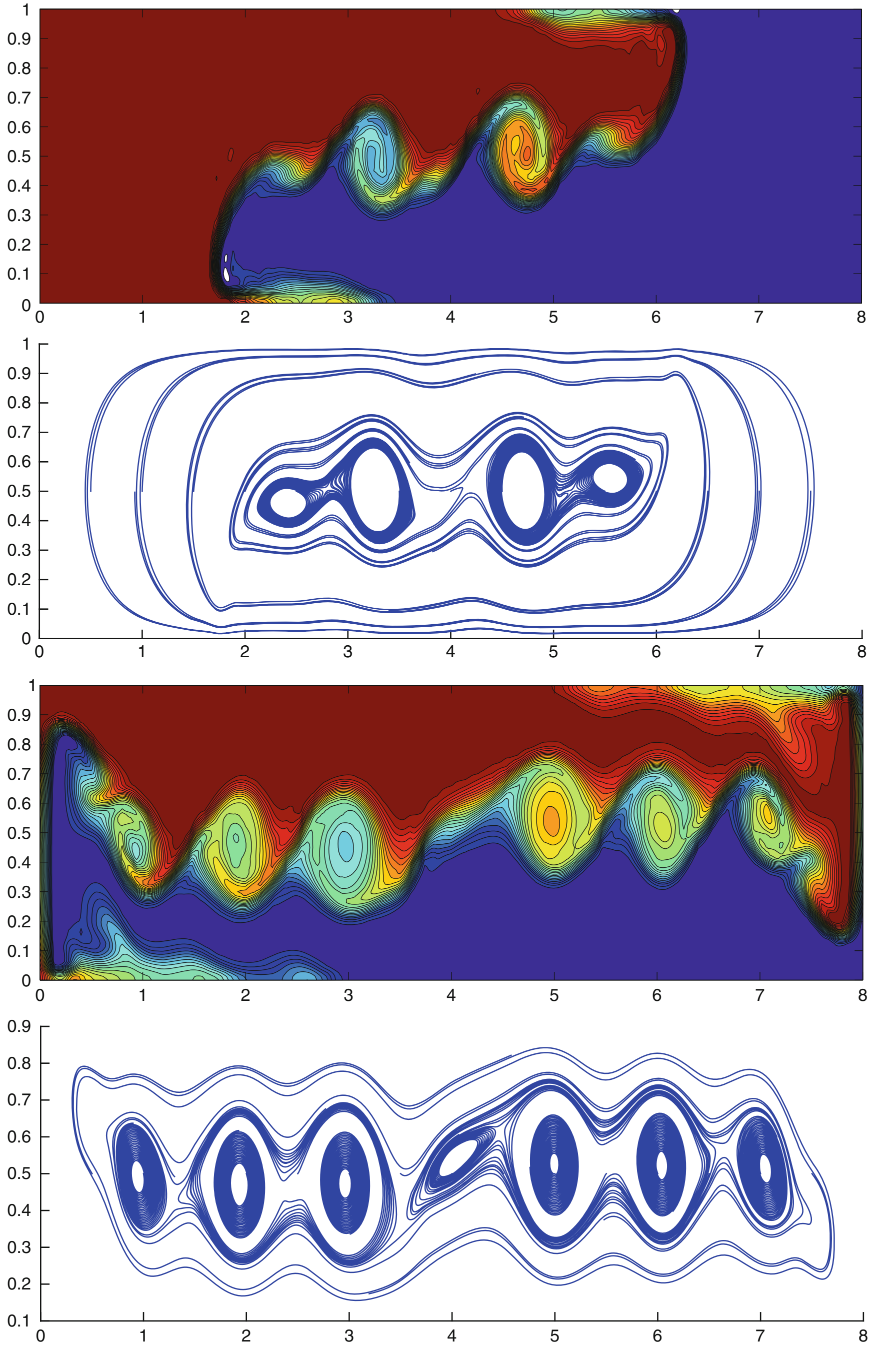


Fig. 2 Temperature contours and velocity streamlines for the $Re = 2{,}000$, $Ri = 4$, $Pr = 1$ Marsigli flow test with $t = 4$, and 8, respectively, with Step 1 and Step 2

4 Conclusions

Our experiments reveal the proposed decoupled VMS method for non-isothermal flows is second order, and performs well on the Marsigli flow test problem. A next step for this method is to perform a full numerical analysis of it, which could reveal further improvements.

Acknowledgements The author L.R. acknowledges partial support from National Science Foundation Grant NSF DMS1522191.

References

1. M. Belenli, S. Kaya, L. Rebholz, An explicitly decoupled variational multiscale method for incompressible, non-isothermal flows. Comput. Methods Appl. Math. **15**, 1–20 (2015)
2. F. Hecht, New development in FreeFem++. J. Numer. Math. **20**, 251–265 (2012)
3. T. Hughes, Multiscale phenomena: green's functions, the Dirichlet-to-Neumann formulation, subgrid-scale models, bubbles and the origin of stabilized methods. Comput. Methods Appl. Mech. Eng. **127**, 387–401 (1995)
4. T. Hughes, L. Mazzei, K. Jansen, Large eddy simulation and variational multiscale method. Comput. Vis. Sci. **3**, 47–59 (2000)
5. T. Hughes, L. Mazzei, A. Oberai, A. Wray, The multiscale formulation of large eddy simulation: decay of homogeneous isotropic turbulence. Phys. Fluids **13**, 505–512 (2001)
6. T. Hughes, A. Oberai, L. Mazzei, Large eddy simulation of turbulent channel flows by the variational multiscale method. Phys. Fluids **13**, 1784–1799 (2001)
7. V. John, On large eddy simulation and variational multiscale methods in the numerical simulation of turbulent incompressible flows. Appl. Math. **51**, 321–353 (2006)
8. V. John, M. Roland, Simulations of the turbulent channel flow at $Re_\tau = 180$ with projection-based finite element variational multiscale methods. Int. J. Numer. Methods Fluids **55**, 407–429 (2007)
9. W. Layton, L. Röhe, H. Tran, Explicitly uncoupled VMS stabilization of fluid flow. Comput. Methods Appl. Mech. **200**, 3183–3199 (2011)
10. J.G. Liu, C. Wang, H. Johnston, A fourth order scheme for incompressible Boussinesq equations. J. Sci. Comput. **18**(2), 253–285 (2003)
11. I. Monteiro, Numerical methods for regularization models for geophysical flows. Universidade Federal do Rio Grande do Sul, Ph.d thesis (2015)

Numerical Experiments for Multiscale Problems in Linear Elasticity

Orane Jecker and Assyr Abdulle

Abstract This paper gives numerical experiments for the Finite Element Heterogeneous Multiscale Method applied to problems in linear elasticity, which has been analyzed in Abdulle (Math Models Methods Appl Sci 16:615–635, 2006). The main results for the FE-HMM a priori errors are stated and their sharpness are verified though numerical experiments.

1 Introduction

Consider the linear elasticity equation in a bounded domain $\Omega \subset \mathbb{R}^d$ with a Lipschitz continuous boundary $\partial\Omega$,

$$\begin{aligned} -\frac{\partial}{\partial x_j}\left(a^{\varepsilon}_{ijkl}\frac{\partial u^{\varepsilon}_k}{\partial x_l}\right) &= f_i, \text{ in } \Omega, \\ u^{\varepsilon}_i &= 0, \text{ on } \partial\Omega, \end{aligned} \tag{1}$$

for $i = 1, \ldots, d$ and where $f \in L^2(\Omega)^d$. Further assume that $a^{\varepsilon}(x)$ is a fourth-order tensor indexed by ε describing the microscopic scale of the problem. We define $|M| = (M : M)^{1/2} = \left(\sum_{i,j=1}^{d} M_{ij}^2\right)^{1/2}$ for any square matrix M. The tensor is such that $a^{\varepsilon}_{ijkl}(x) \in L^{\infty}(\Omega)$, for all $i, j, k, l = 1, \ldots, d$, and

$$a^{\varepsilon}_{ijkl} = a^{\varepsilon}_{jikl} = a^{\varepsilon}_{klij}, \tag{2}$$

$$\alpha|M|^2 \leq a^{\varepsilon}M : M, \text{ for any symmetric matrix } M, \tag{3}$$

$$|a^{\varepsilon}M| \leq \beta|M|, \text{ for any symmetric matrix } M, \tag{4}$$

O. Jecker (✉) • A. Abdulle
École Polytechnique Fédérale de Lausanne, ANMC, CH-1015 Lausanne, Switzerland
e-mail: orane.jecker@epfl.ch; assyr.abdulle@epfl.ch

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_13

where $0 < \alpha \leq \beta < \infty$. We define the linearized strain tensor e, for $i,j = 1,\dots,d$, by

$$e(u^\varepsilon) = (e_{ij}(u^\varepsilon))_{1\leq i,j\leq d}, \quad e_{ij}(u^\varepsilon) = \frac{1}{2}\left(\frac{\partial u_i^\varepsilon}{\partial x_j} + \frac{\partial u_j^\varepsilon}{\partial x_i}\right).$$

The weak formulation of problem (1) reads: find $u^\varepsilon \in H_0^1(\Omega)^d$ such that

$$B(u^\varepsilon, v) := \int_\Omega a^\varepsilon(x)e(u^\varepsilon) : e(v)\mathrm{d}x = \int_\Omega fv\mathrm{d}x =: F(v), \tag{5}$$

for all $v \in H_0^1(\Omega)^d$. Problem (5) is well-posed thanks to the first Korn inequality, that is

$$\|v\|_{\mathrm{H}^1(\Omega)} \leq C\left(\int_\Omega |e(v)|^2\mathrm{d}x\right)^{1/2}.$$

Solving (5) with standard FEM requires the mesh size to be smaller than the fine scale, which is prohibitive if ε is small. However, the effective dynamics of the problem can be described using homogenization theory [6, 9]. Using the theory of H-convergence [5, 8], it can be established that a subsequence of the family of solutions $\{u^\varepsilon\}$ converges weakly to an effective solution u^0, satisfying the homogenized formulation

$$B_0(u^0, v) := \int_\Omega a^0(x)e(u^0) : e(v)\mathrm{d}x = F(v), \quad \forall v \in H_0^1(\Omega)^d. \tag{6}$$

The homogenized tensor a^0 verifies the properties (2), (3), and (4) for some constants $0 < \alpha_0 \leq \beta_0 < \infty$. Under additional information on the small scale of the tensor, such as periodicity

(H1) $a^\varepsilon(x) = a(x/\varepsilon) = a(y)$ is Y-periodic in y, where $Y = (0,1)^d$,

explicit equations are available to compute the homogenized tensor a^0

$$a^0_{ijkl} = \frac{1}{|Y|}\int_Y a_{ijkl}(y) + \sum_{h,m=1}^d a_{ijhm}(y)\frac{\partial \chi_h^{kl}(y)}{\partial y_m}\mathrm{d}y.$$

The functions $\chi_h^{kl} \in W_{per}(Y)$ are solutions of the micro problems

$$-\frac{\partial}{\partial y_j}\left(a_{ijhm}\frac{\partial \chi_h^{kl}}{\partial y_m}\right) = \frac{\partial a_{ijkl}}{\partial y_j}, \text{ in } Y, \text{ for } i = 1,\dots,d, \tag{7}$$

with periodic boundary conditions. The space $W_{\text{per}}(Y)$ is defined as

$$W_{\text{per}}(Y) = \{v \in H^1_{per}(Y)^d \mid \int_Y v_i \mathrm{d}y = 0, i = 1, \dots, d\}.$$

Remark Problem (1) can be easily adapted to non-homogeneous Dirichlet and Neumann boundary conditions. A lifting of the Dirichlet data should be considered and extra terms are added to the weak formulations (5) and (6).

In Sect. 2, we state the FE-HMM method for linear elasticity [1] and in Sect. 3 we recall the a priori error estimates derived in [1, 2]. Finally, in Sect. 4, we illustrate the sharpness of the convergence rates with numerical examples.

2 Finite Element Heterogeneous Multiscale Method for Linear Elasticity

The FE-HMM gives us a macroscopic solution based on a macro to micro modeling without knowing the homogenized tensor a^0.

Macro Problem Let $\mathscr{T}_H$ be a mesh over Ω with mesh size $H \gg \varepsilon$ given by $H = \max_{K \in \mathscr{T}_H} h_K$. In each macro element K, we consider integration nodes $x_{j,K}$ and weights $\omega_{j,K}$, for $j = 1, \dots, J$, and construct sampling domains $K_{\delta_j} = x_{j,K} + \delta[-1/2, 1/2]^d$. We define a macro FE space of degree p by

$$V^p(\Omega, \mathscr{T}_H) = \{v^H \in H^1_0(\Omega)^d \mid v^H|_K \in \mathscr{R}^p(K)^d, \quad \forall K \in \mathscr{T}_H\},$$

where $\mathscr{R}^p(K)$ is the space $\mathscr{P}^p(K)$ of polynomials on K of degree at most p if K is a triangle, or the space $\mathscr{Q}^p(K)$ of polynomials on K of degree at most p in each variables if K is a rectangle. We construct a macro bilinear form

$$B_H(v^H, w^H) := \sum_{K \in \mathscr{T}_H} \sum_{j=1}^{J} \frac{\omega_{j,K}}{|K_{\delta_j}|} \int_{K_{\delta_j}} a^\varepsilon(x) e(v_j^h) : e(w_j^h) \mathrm{d}x,$$

where v_j^h (resp. w_j^h) is the solution of the micro problem (9) on the sampling domain K_{δ_j}. The FE-HMM solution u^H verifies

$$B_H(u^H, v^H) = F(v^H), \quad \forall v^H \in V^p(\Omega, \mathscr{T}_H). \tag{8}$$

Micro Problem Let $\mathscr{T}_h$ be a micro partition over K_{δ_j}, for $j = 1, \dots, J$, of mesh size $h << \varepsilon$, with $h = \max_{K \in \mathscr{T}_h} h_K$. For each K_{δ_j}, we define a micro FE space of degree q as

$$S^q(K_{\delta_j}, \mathscr{T}_h) = \{v^h \in W(K_{\delta_j}) \mid v^h|_K \in \mathscr{R}^q(K)^d, \quad \forall K \in \mathscr{T}_h\}.$$

The micro problems read: find u_j^h such that $(u_j^h - u_{lin,j}^H) \in S^q(K_{\delta_j}, \mathcal{T}_h)$ and

$$\int_{K_{\delta_j}} a^\varepsilon(x) e(u_j^h) : e(v_j^h) \mathrm{d}x = 0, \quad \forall v_j^h \in S^q(K_{\delta_j}, \mathcal{T}_h), \tag{9}$$

where $u_{lin,j}^H(x) = u^H(x_{j,K}) + (x - x_{j,K}) e(u^H(x_{j,K}))$ is a linearization of u^H taken at the quadrature node $x_{j,K}$. The space $W(K_{\delta_j})$ sets the coupling between the micro and macro solvers and depends on the choice of boundary conditions in problem (9),

$$W(K_{\delta_j}) = W_{\text{per}}(K_{\delta_j}) \text{ for periodic coupling, or ,}$$
$$W(K_{\delta_j}) = W_{\text{dir}}(K_{\delta_j}) = H_0^1(K_{\delta_j})^d \text{ for Dirichlet coupling.}$$

3 A Priori Error Estimates

In this section we give a priori error estimates for the FE-HMM method, details can be found in [1, 2]. The error is decomposed into the macro, modeling, and micro error,

$$\|u^0 - u^H\| \le e_{\text{MAC}} + e_{\text{MOD}} + e_{\text{MIC}}.$$

We assume that the micro solution χ^{lm} (solution of Eq. (7)) are smooth enough, i.e.,

(H2) $\varepsilon\chi^{lm} \in H^{q+1}(K_{\delta_j})^d$ with $\|D^\alpha(\varepsilon\chi^{lm})\|_{\mathrm{L}^\infty(K_{\delta_j})} \le C\varepsilon^{-|\alpha|+1}$, for $\alpha \le q+1$ and $l, m = 1, \ldots, d$.

Theorem 1 ([1]) *Let u^0 and u^H be solutions of* (6) *and* (8)*, respectively. Assume that $u^0 \in H^{r+1}(\Omega)^d$, for some $r > 0$, and that the hypothesis (H2) holds. Then,*

$$\|u^0 - u^H\|_{\mathrm{H}^1(\Omega)} \le C\left(H^s + \left(\frac{h}{\varepsilon}\right)^{2q} + e_{MOD}\right),$$

$$\|u^0 - u^H\|_{\mathrm{L}^2(\Omega)} \le C\left(H^{s+1} + \left(\frac{h}{\varepsilon}\right)^{2q} + e_{MOD}\right), \quad s = \min(r, p).$$

If in addition, the hypothesis (H1) holds, the modeling error is given by

$$e_{MOD} = 0, \text{ for periodic coupling with } \delta/\varepsilon \in \mathbb{N}^*,$$
$$e_{MOD} = \frac{\varepsilon}{\delta}, \text{ for Dirichlet coupling with } \delta > \varepsilon.$$

The homogeneous tensor can be approximated during the assembling process of the FE-HMM. For general tensors and sampling domains, we have, in each macro element $K \in \mathcal{T}_H$,

$$a^{0,h}_{ijkl}(x_{m,K}) = \frac{1}{|K_{\delta_m}|} \int_{K_{\delta_m}} a^{\varepsilon}(x) e(\varphi^h_{m,i,j}) : e(\varphi^h_{m,k,l}) \mathrm{d}x, \tag{10}$$

where $x_{m,K}$ is a quadrature point in K, and K_{δ_m} is the sampling domain around $x_{m,K}$. The functions $\varphi^h_{m,i,j} \in W(K_{\delta_m})$ are solutions of (9) for $i, j \in \{1, \dots, d\}$. Then, note that if (H1) holds, the tensors a^0 and $a^{0,h}$ are constants in Ω. The error introduced by computing a^0 is given by the following Lemma.

Lemma 2 *Assume that (H1) holds and that periodic coupling is used with $\delta/\varepsilon \in \mathbb{N}^*$. Let $a^{0,h} = (a^{0,h}_{ijkl})$ be defined in (10). It holds*

$$|a^0_{ijkl} - a^{0,h}_{ijkl}| \le C \left(\frac{h}{\varepsilon}\right)^{2q}.$$

Proof Follows from [1] (see also [3, 4]).

4 Numerical Experiments

In this section we present numerical examples to verify the sharpness of the bounds obtained in Theorem 1 and Lemma 2. In Table 1, we show the best refinement strategies for the optimal H^1 and L^2 convergence rates with minimal computational cost.

We start by showing that the macro convergence rates in H are sharp. Let $\varepsilon = 1/10$, and consider Eq. (1) in $\Omega = [0,1]^2$ with homogeneous Dirichlet boundary condition, a right-hand side $f = 1$, and a tensor $a^{\varepsilon}(x) = a(x/\varepsilon) = a(y)$ given by

$$a(y) = \begin{pmatrix} \sin(2\pi y_1) + 2 & 0 & 0 \\ 0 & \sin(2\pi y_2) + 2 & 0 \\ 0 & 0 & 10 \end{pmatrix}, \quad a^0 = \begin{pmatrix} \sqrt{3} & 0 & 0 \\ 0 & \sqrt{3} & 0 \\ 0 & 0 & 10 \end{pmatrix}.$$

Table 1 Best refinement strategies for optimal convergence rates

Macro FE	u^0	Micro FE	H^1 norm	L^2 norm
P^1	$H^2(\Omega)$	P^1	$\sqrt{N_{mac}} = N_{mic}$	$N_{mac} = N_{mic}$
		P^2	$N_{mac}^{1/4} = N_{mic}$	$\sqrt{N_{mac}} = N_{mic}$
P^2	$H^3(\Omega)$	P^1	$N_{mac} = N_{mic}$	$N_{mac}^{3/2} = N_{mic}$
		P^2	$\sqrt{N_{mac}} = N_{mic}$	$N_{mac}^{3/4} = N_{mic}$

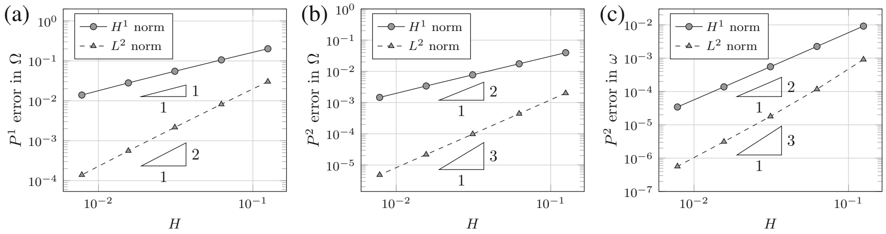


Fig. 1 Error between u^0 (solution of (6)) and u^H (solution of (8)) in Ω for (**a**) P^1 macro and micro FE spaces and (**b**) P^2 macro and micro FE spaces. The setting in (**c**) is similar to (**b**) but with $\omega \subset \Omega$

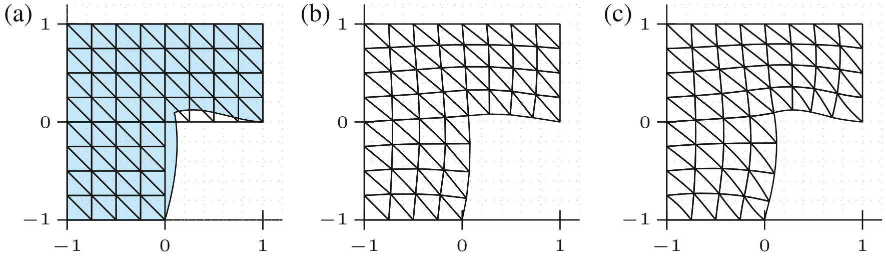


Fig. 2 (**a**) Reference solution. Finite element solution u^H for P^1 macro and micro FE (**b**), and P^2 macro and micro FE (**c**)

A reference solution for u^0 is computed on a very fine mesh. We use periodic coupling with $\delta = \varepsilon$ in order to have zero modeling error. Further, the micro degrees of freedom is chosen such that the micro error can be neglected, and take $H = 1/8, 1/16, 1/32, 1/64, 1/128$. In Fig. 1a, we monitor the H^1 and L^2 errors to u^0 for the piecewise macro and micro FE-HMM. The solution u^0 is in $H^2(\Omega)$ and one can see the linear and quadratic rates for the piecewise H^1 and L^2 errors, respectively. However, as one can see in Fig. 1b, u^0 is not smooth enough to observe the H^2 and H^3 convergence rates for the quadratic H^1 and L^2 norms, respectively. The optimal rates can be seen in Fig. 1c where we restrict the errors to a subdomain $\omega \subset \Omega$ to avoid corner singularities.

Consider now problem (1) with $f = 1$, on a L-shaped domain centered around $(0, 0)$ with width 2. We impose free Neumann boundary condition on $\{x = 0, y \in [-1, 0]\}$ and $\{y = 0, x \in [0, 1]\}$, and homogeneous Dirichlet boundary condition elsewhere.

In Fig. 2a one can see the reference displacement in comparison to the initial coarse mesh. Using periodic coupling and $\delta = \varepsilon$, we compute the FE-HMM solutions for P^1 macro and micro FE and for P^2 macro and micro FE; they are shown in Fig. 2b, c, respectively.

In Fig. 3a, b, we plot the H^1 and L^2 convergence rates for P^1 macro and micro FE spaces. The optimal refinement follows the ratio given in Table 1.

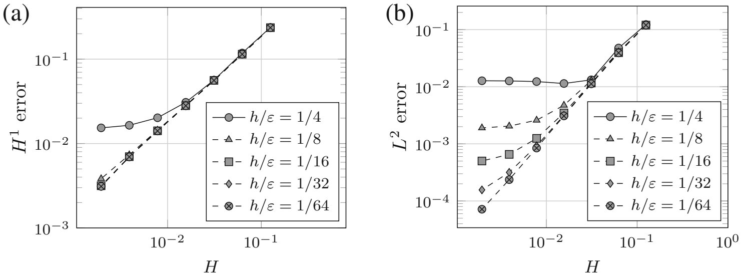


Fig. 3 H^1 (**a**) and L^2 (**b**) errors between u^0 and u^H for piecewise macro and micro FE spaces

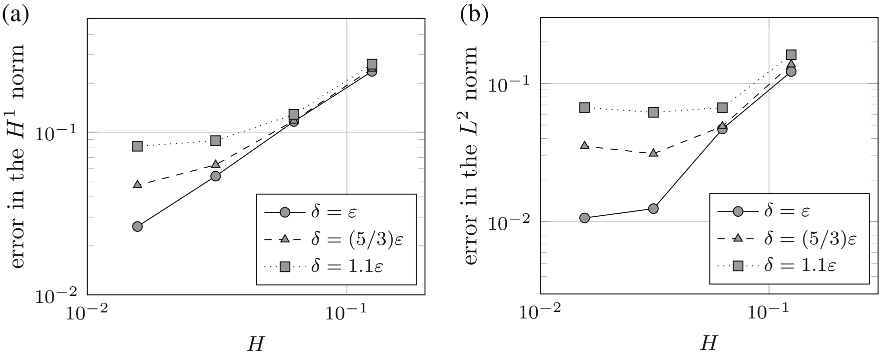
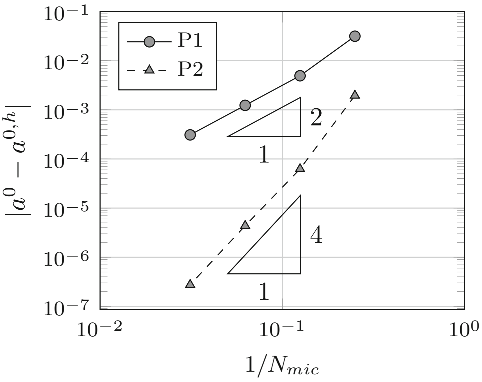


Fig. 4 H^1 error (**a**) and L^2 error (**b**) between the homogenized solution and the FE-HMM with Dirichlet coupling for $\delta = 5/3\varepsilon$ (*dashed*) and $\delta = 1.1\varepsilon$ (*dotted*). The error $\delta = \varepsilon$ (*full*) is obtained with periodic coupling

We show next the influence of the modeling error on the same problem with sampling domains K_δ with $\delta > \varepsilon$. We take $H = 1/8, 1/16, 1/32$, and $1/64$, with micro mesh size sufficiently small to eliminate the micro error. We use piecewise FE for the macro and micro problems. The size of the sampling domains K_δ are $\delta = 5/3\varepsilon$ and $\delta = 1.1\varepsilon$, and for those values we solve the micro problems (9) with homogeneous Dirichlet boundary conditions. In Fig. 4a, b, we see that the choice of δ has an important influence in the error. Increasing the size of the sampling domain from $\delta = 1.1\varepsilon$ to $\delta = 5/3\varepsilon$ improves the quality of the error, as expected from Theorem 1. The periodic coupling with $\delta = \varepsilon$ gives the optimal convergence rate since the modeling error is zero, as predicted by Theorem 1.

Modeling error and random coefficients The use of artificial boundary conditions for the micro problem (9) leads to a modeling (or resonance) error of size $\mathcal{O}(\varepsilon/\delta)$ for elliptic problems. Such error terms also appear for problems with random stationary fields, where (9) is usually defined in the whole $\mathbb{R}^d$ [10]. Truncations using either Dirichlet or periodic boundary conditions can then be used for numerical approximation. In [7], a reduction of this resonance error is obtained by adding a zero-order term the cell problem (9) and using a suitable Richardson extrapolation

Fig. 5 Convergence rates $|a^0 - a^{0,h}|$ with respect to N_{mic}^{-1} for P^1(*full*) and P^2(*dashed*) micro FE spaces



of the modified cell problem. Such strategies could also be of interest for elastic problems.

Finally, we study the bound in Lemma 2. We use piecewise FE for the macro problem and compare the exact homogenized tensor with the numerical homogenized tensor. In Fig. 5, we show the convergence rate

$$|a^0_{1111} - a^{0,h}_{1111}| = |\sqrt{3} - a^{0,h}_{1111}|,$$

for piecewise (full lines) and quadratic (dashed lines) micro FE, and observe the expected rates.

Acknowledgements This work was supported in part by the Swiss National Science Foundation.

References

1. A. Abdulle, Analysis of a heterogeneous multiscale FEM for problems in elasticity. Math. Models Methods Appl. Sci. **16**(4), 615–635 (2006)
2. A. Abdulle, The finite element heterogeneous multiscale method: a computational strategy for multiscale PDEs, in *Multiple Scales Problems in Biomathematics, Mechanics, Physics and Numerics*. GAKUTO International Series. Mathematical Sciences and Applications, vol. 31 (Gakkōtosho, Tokyo, 2009), pp. 133–181
3. A. Abdulle, A priori and a posteriori error analysis for numerical homogenization: a unified framework. Ser. Contemp. Appl. Math. CAM **16**, 280–305 (2011)
4. A. Abdulle, Discontinuous Galerkin finite element heterogeneous multiscale method for elliptic problems with multiple scales. Math. Comput. **81**(278), 687–713 (2012)
5. G. Allaire, *Shape Optimization by the Homogenization Method*. Applied Mathematical Sciences, vol. 146 (Springer, New York, 2002)
6. D. Cioranescu, P. Donato, *An introduction to homogenization*. Oxford Lecture Series in Mathematics and its Applications, vol. 17 (Oxford University Press, New York, 1999)
7. A. Gloria, Z. Habibi, Reduction of the resonance error – part 2: approximation of correctors, extrapolation, and spectral theory, Preprint, hal-00933234 (2014)

8. F. Murat, L. Tartar, *H*-convergence, in *Topics in the Mathematical Modelling of Composite Materials*. Progress in Nonlinear Differential Equations and Their Applications, vol. 31 (Birkhäuser Boston, Boston, 1997), pp. 21–43
9. O.A. Oleinik, A. Shamaev, G. Yosifian, *Mathematical problems in elasticity and homogenization* (North-Holland, Amsterdam, 1992)
10. G.C. Papanicolaou, S.R.S. Varadhan, Boundary value problems with rapidly oscillating random coefficients, in *Random Fields, vol. I, II (Esztergom, 1979)*. Colloq. Math. Soc. János Bolyai, vol. 27 (North-Holland, Amsterdam/New York, 1981), pp. 835–873

The Skeleton Reduction for Finite Element Substructuring Methods

Christian Wieners

Abstract We introduce an abstract concept for decomposing spaces with respect to a substructuring of a bounded domain. In this setting we define weakly conforming finite element approximations of quadratic minimization problems. Within a saddle point approach the reduction to symmetric positive Schur complement systems on the skeleton is analyzed. Applications include weakly conforming variants of least squares and minimal residuals.

We consider general weakly conforming substructuring methods and its hybridization for the approximation of linear differential equations on Lipschitz domains $\Omega \subset \mathbb{R}^D$. The discretization is based on a decomposition $\Omega_h = \bigcup_{K\in\mathscr{K}} K$ into convex open subdomains $K \subset \Omega$ with weak continuity constraints on the skeleton $\Gamma = \bigcup \partial K = \overline{\Omega} \setminus \Omega_h$. Here we present a general concept for the analysis of such discretizations based on corresponding saddle point formulations, and following [7] we consider the reduction to degrees of freedom to the skeleton. For comparison, we also summarize the DPG method [4] in this setting using formal trace mappings arising from integration by parts and quotient spaces replacing trace spaces.

1 Substructuring, Trace Spaces, and Minimization

Let L be a linear first-order differential operator with $Lv \in \mathrm{L}_2(\Omega, \mathbb{R}^M)$ for $v \in \mathrm{C}^1_0(\Omega, \mathbb{R}^N)$, and let L^{ad} be its adjoint operator with

$$(Lv, w)_\Omega = (v, L^{\mathrm{ad}} w)_\Omega\,, \quad v \in \mathrm{C}^1_0(\Omega, \mathbb{R}^N)\,,\ w \in \mathrm{C}^1_0(\Omega, \mathbb{R}^M)\,.$$

C. Wieners (✉)
Institut für Angewandte und Numerische Mathematik, KIT, Karlsruhe, Germany
e-mail: christian.wieners@kit.edu

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_14

We define for L in Ω (and analogously for L^{ad} and for open subsets of Ω)

$$\mathrm{H}(L,\Omega) = \big\{v \in \mathrm{L}_2(\Omega,\mathbb{R}^N)\colon f \in \mathrm{L}_2(\Omega,\mathbb{R}^M) \text{ exists with } (f,w)_\Omega = (v, L^{\mathrm{ad}}w)_\Omega \text{ for all } w \in \mathrm{C}^1_0(\Omega,\mathbb{R}^M)\big\}\,.$$

Then, L extends to this space, and $\mathrm{H}(L,\Omega)$ is a Hilbert spaces with respect to the graph norm $\|v\|_{L,\Omega} = \sqrt{\|v\|^2_\Omega + \|Lv\|^2_\Omega}$.

For open subsets $K \subset \Omega$ we define the bilinear map

$$\gamma_K(v,w) = (Lv,w)_K - (v,L^{\mathrm{ad}}w)_K\,, \qquad v \in \mathrm{H}(L,K)\,,\ w \in \mathrm{H}(L^{\mathrm{ad}},K)$$

and the kernels

$$\mathrm{H}_0(L,K) = \big\{v \in \mathrm{H}(L,K)\colon \gamma_K(v,w) = 0 \text{ for all } w \in \mathrm{H}(L^{\mathrm{ad}},K)\big\}\,,$$
$$\mathrm{H}_0(L^{\mathrm{ad}},K) = \big\{w \in \mathrm{H}(L^{\mathrm{ad}},K)\colon \gamma_K(v,w) = 0 \text{ for all } v \in \mathrm{H}(L,K)\big\}\,.$$

By definition, we have $\gamma_K(v,w) = \gamma_K(v+v_0,w+w_0)$ for $v \in \mathrm{H}(L,K)$, $v_0 \in \mathrm{H}_0(L,K)$, $w \in \mathrm{H}(L^{\mathrm{ad}},K)$, $w_0 \in \mathrm{H}_0(L^{\mathrm{ad}},K)$, so that $\gamma_K(\cdot,\cdot)$ extends to the quotient spaces

$$\hat{\mathrm{H}}(L,K) = \mathrm{H}(L,K)/\mathrm{H}_0(L,K)\,, \qquad \hat{\mathrm{H}}(L^{\mathrm{ad}},K) = \mathrm{H}(L^{\mathrm{ad}},K)/\mathrm{H}_0(L^{\mathrm{ad}},K)$$

with equivalence classes $\hat{v} = v + \mathrm{H}_0(L,K)$, $\hat{w} = w + \mathrm{H}_0(L^{\mathrm{ad}},K)$ and norms

$$\|\hat{v}\|_{L,\partial K} = \inf_{\hat{v}=v+\mathrm{H}_0(L,K)} \|v\|_{L,K}\,, \qquad \|\hat{w}\|_{L^{\mathrm{ad}},\partial K} = \inf_{\hat{w}=w+\mathrm{H}_0(L^{\mathrm{ad}},K)} \|w\|_{L^{\mathrm{ad}},K}\,.$$

By construction, $\gamma_K(\cdot,\cdot)$ is continuous, i.e.,

$$|\gamma_K(v,w)| \le \|\hat{v}\|_{L,\partial K}\|\hat{w}\|_{L^{\mathrm{ad}},\partial K} \le \|v\|_{L,K}\|w\|_{L^{\mathrm{ad}},K}$$

for $v \in \mathrm{H}(L,K)$, $w \in \mathrm{H}(L^{\mathrm{ad}},K)$, $\hat{v} = v + \mathrm{H}_0(L,K)$, $\hat{w} = v + \mathrm{H}_0(L^{\mathrm{ad}},K)$.

Lemma 1 (cf. Lem. 2.2 in [3]) *We have*

$$\|\hat{v}\|_{L,\partial K} = \sup_{\hat{w}\in\hat{\mathrm{H}}(L^{\mathrm{ad}},K)} \frac{\gamma_K(\hat{v},\hat{w})}{\|\hat{w}\|_{L^{\mathrm{ad}},\partial K}}\,, \qquad \|\hat{w}\|_{L^{\mathrm{ad}},\partial K} = \sup_{\hat{v}\in\hat{\mathrm{H}}(L,K)} \frac{\gamma_K(\hat{v},\hat{w})}{\|\hat{v}\|_{L,\partial K}}\,.$$

Proof For given $\hat{v} \in \hat{\mathrm{H}}(L,K)$ define $w_{\hat{v}} \in \mathrm{H}(L^{\mathrm{ad}},K)$ solving

$$(w_{\hat{v}},w)_{L^{\mathrm{ad}},K} = \gamma_K(\hat{v},w)\,, \qquad w \in \mathrm{H}(L^{\mathrm{ad}},K)\,. \tag{1}$$

Then we set $v_{\hat{v}} = -L^{\mathrm{ad}} w_{\hat{v}}$, and (1) yields for $w \in \mathrm{C}^1_0(K, \mathbb{R}^M)$

$$0 = (w_{\hat{v}}, w)_{L^{\mathrm{ad}},K} = (w_{\hat{v}}, w)_K - (v_{\hat{v}}, L^{\mathrm{ad}} w)_K \,.$$

Thus, $v_{\hat{v}} \in \mathrm{H}(L, K)$, $Lv_{\hat{v}} = w_{\hat{v}}$, and $\|v_{\hat{v}}\|_{L,K} = \|w_{\hat{v}}\|_{L^{\mathrm{ad}},K}$. Moreover,

$$\gamma_K(\hat{v}, w) = (w_{\hat{v}}, w)_{L^{\mathrm{ad}},K} = (Lv_{\hat{v}}, w)_K - (v_{\hat{v}}, L^{\mathrm{ad}} w)_K = \gamma_K(v_{\hat{v}}, w)$$

for $w \in \mathrm{H}_0(L^{\mathrm{ad}}, K)$, i.e., $\hat{v} = v_{\hat{v}} + \mathrm{H}_0(L, K)$. This finally yields

$$\begin{aligned}\|\hat{v}\|_{L,\partial K} \le \|v_{\hat{v}}\|_{L,K} = \|w_{\hat{v}}\|_{L^{\mathrm{ad}},K} &= \frac{\gamma_K(\hat{v}, w_{\hat{v}})}{\|w_{\hat{v}}\|_{L^{\mathrm{ad}},K}} \\ &\le \sup_{w \in \mathrm{H}(L^{\mathrm{ad}},K)} \frac{\gamma_K(\hat{v}, w)}{\|w\|_{L^{\mathrm{ad}},K}} = \sup_{\hat{w} \in \hat{\mathrm{H}}(L^{\mathrm{ad}},K)} \frac{\gamma_K(\hat{v}, \hat{w})}{\|\hat{w}\|_{L^{\mathrm{ad}},\partial K}} \le \|\hat{v}\|_{L,\partial K} \,. \qquad \square\end{aligned}$$

On the broken spaces $\mathrm{H}(L, \Omega_h) = \prod_K \mathrm{H}(L, K)$ and $\mathrm{H}(L^{\mathrm{ad}}, \Omega_h)$ we define

$$\gamma_h(v, w) = \sum_K \gamma_K(v_K, w_K)\,, \qquad v \in \mathrm{H}(L, \Omega_h)\,,\ w \in \mathrm{H}(L^{\mathrm{ad}}, \Omega_h)\,,$$

where we use the notation $v_K = v|_K$ and $w_K = w|_K$. Again, $\gamma_h(\cdot, \cdot)$ extends to the quotient spaces $\hat{\mathrm{H}}(L, \Omega_h) = \prod_K \hat{\mathrm{H}}(L, K)$ and $\hat{\mathrm{H}}(L^{\mathrm{ad}}, \Omega_h)$.

Boundary conditions In many cases, the space $\mathrm{H}_0(L, \Omega)$ is too small, but L is not injective on $\mathrm{H}(L, \Omega)$. Thus we select a subspace $V \subset \mathrm{H}(L, \Omega)$ with $\mathrm{H}_0(L, \Omega) \subset V$ such that $C_V > 0$ exists satisfying

$$\|v\|_\Omega \le C_V \, \|Lv\|_\Omega \,, \qquad v \in V \,, \tag{2}$$

see [7, Sect.2] for various examples. The adjoint space is given by

$$W = \big\{ w \in \mathrm{L}_2(\Omega, \mathbb{R}^M) \colon \exists g \in \mathrm{L}_2(\Omega, \mathbb{R}^N) \text{ with } (Lv, w)_\Omega = (v, g)_\Omega \,,\ v \in V \big\} \,.$$

Then, $W \subset \mathrm{H}(L^{\mathrm{ad}}, \Omega)$, and we observe

$$\begin{aligned} V &= \big\{ v \in \mathrm{H}(L, \Omega_h) \colon \gamma_h(v, w) = 0 \text{ for all } w \in W \big\} \,, \\ W &= \big\{ w \in \mathrm{H}(L^{\mathrm{ad}}, \Omega_h) \colon \gamma_h(v, w) = 0 \text{ for all } v \in V \big\} \,, \end{aligned}$$

cf. [5] and Thm. 2.1 in [2]. The corresponding trace spaces are

$$\begin{aligned} \hat{V} &= \big\{ \big(v_K + \mathrm{H}_0(L, K)\big)_{K \in \mathscr{K}} \in \hat{\mathrm{H}}(L, \Omega_h) \colon v \in V \big\} \,, \\ \hat{W} &= \big\{ \big(w_K + \mathrm{H}_0(L^{\mathrm{ad}}, K)\big)_{K \in \mathscr{K}} \in \hat{\mathrm{H}}(L^{\mathrm{ad}}, \Omega_h) \colon w \in W \big\} \,. \end{aligned}$$

Minimization On the broken space we consider the quadratic functional

$$J_h(v) = \frac{1}{2} a_h(v, v) - (f, v)_\Omega \,, \qquad v \in \mathrm{H}(L, \Omega_h) \tag{3}$$

with a symmetric bilinear form $a_h(v, v) = \sum_K a_K(v_K, v_K)$ satisfying

$$\alpha \, \|v\|^2_{L,\Omega} \le a_h(v, v) \le C_a \|v\|^2_{L,\Omega_h} \,, \qquad v \in V \,,$$

and $f \in L_2(\Omega, \mathbb{R}^N)$. Then, a unique minimizer $u \in V$ of $J(\cdot) = J_h(\cdot)|_V$ exists characterized by $a_h(u, v) = (f, v)_\Omega$ for $v \in V$, and we directly obtain:

Lemma 2

(a) Let $(u, \hat\mu) \in \mathrm{H}(L, \Omega_h) \times \hat W$ *be a saddle point of*

$$F_h(v, \hat w) = J_h(v) + \gamma_h(v, \hat w) \,, \qquad (v, \hat w) \in \mathrm{H}(L, \Omega_h) \times \hat W \,.$$

Then, we have $u \in V$, *and* u *is a minimizer of* $J(\cdot)$.

(b) Let $(u, \hat\mu, \hat u) \in \mathrm{H}(L, \Omega_h) \times \hat{\mathrm{H}}(L^{\mathrm{ad}}, \Omega_h) \times \hat V$ *be a saddle point of*

$$\hat F_h(v, \hat w, \hat v) = J_h(v) + \gamma_h(v - \hat v, \hat w) \,, \ (v, \hat w, \hat v) \in \mathrm{H}(L, \Omega_h) \times \hat{\mathrm{H}}(L^{\mathrm{ad}}, \Omega_h) \times \hat V \,.$$

Then, $(u, \hat\mu) \in \mathrm{V} \times \hat W$ *is a saddle point of* $F_h(\cdot, \cdot)$, *and*

$$\hat u = \big(u_K + \mathrm{H}_0(L, K)\big)_{K \in \mathscr{K}} \,.$$

This applies to $a_h(v, v) = (Lv, Lv)_{\Omega_h}$ with $\alpha = (1 + C_V^2)^{-1}$ and $C_a = 1$.

2 Weakly Conforming Approximation

We select discrete spaces $V_K \subset \mathrm{H}(L, K)$ and $W_K \subset \mathrm{H}(L^{\mathrm{ad}}, K)$ for all K, and on Ω_h we define the broken discrete spaces $V_{\mathscr{K}} = \prod V_K \subset \mathrm{H}(L, \Omega_h)$ and $W_{\mathscr{K}} = \prod W_K \subset \mathrm{H}(L^{\mathrm{ad}}, \Omega_h)$. The conforming space $W_h = W_{\mathscr{K}} \cap W$ defines the weakly conforming approximation space

$$V_h = \big\{ v_h \in V_{\mathscr{K}} \colon \gamma_h(v_h, w_h) = 0 \text{ for all } w_h \in W_h \big\} \,.$$

We define $V_h(K) = V_h|_K \subset V_K$, $W_h(K) = W_h|_K \subset W_K$, the kernel spaces

$$\begin{aligned} V_{0,h,K} &= \big\{ v_K \in V_h(K) \colon \gamma_K(v_K, w_K) = 0 \text{ for all } w_K \in W_h(K) \big\} \,, \\ W_{0,h,K} &= \big\{ w_K \in W_h(K) \colon \gamma_K(v_K, w_K) = 0 \text{ for all } v_K \in V_h(K) \big\} \,, \\ W_{0,\mathscr{K}} &= \big\{ w_h \in W_{\mathscr{K}} \colon \gamma_h(v_h, w_h) = 0 \text{ for all } v_h \in V_{\mathscr{K}} \big\} \,, \end{aligned}$$

and the quotient spaces $\hat{V}_h = \prod V_h(K)/V_{0,h,K}$, $\hat{W}_h = \prod W_h(K)/W_{0,h,K}$, and $\hat{W}_{\mathscr{K}} = W_{\mathscr{K}}/W_{0,\mathscr{K}}$. Then, we obtain

$$V_h = \left\{v_h \in V_{\mathscr{K}} \colon \gamma_h(v_h, \hat{w}_h) = 0 \text{ for all } \hat{w}_h \in \hat{W}_h\right\},$$
$$\hat{W}_h = \left\{\hat{w}_h \in \hat{W}_{\mathscr{K}} \colon \gamma_h(\hat{v}_h, \hat{w}_h) = 0 \text{ for all } \hat{v}_h \in \hat{V}_h\right\}.$$

We assume that the quadratic form $a_h(\cdot,\cdot)$ is also uniformly convex in V_h. In addition, inf-sup stability is required. Therefore, the selection of the discrete spaces V_K, W_K has to be well balanced such that $\alpha_0, \beta_0 > 0$ exists with

$$a_h(v_h, v_h) \geq \alpha_0 \|v_h\|^2_{L,\Omega_h}, \qquad \sup_{v_K \in V_K} \frac{\gamma_K(v_K, \hat{w}_K)}{\|v_K\|_{L,K}} \geq \beta_0 \|\hat{w}_K\|_{L^{\mathrm{ad}},\partial K} \tag{4}$$

for all $v_h \in V_h$ and $w_K \in W_K$. Note that we have $\alpha_0 \leq \alpha$ and $\beta_0 \leq 1$.

From saddle point theory [1, Prop. 2.5, 2.5, 2.7] we obtain:

Theorem 3

(a) A unique minimizer $u_h \in V_h$ of $J_h(\cdot)$ exists solving $a_h(u_h, v_h) = (f, v_h)_\Omega$ for $v_h \in V_h$. The error is bounded by

$$\|u - u_h\|_{L,\Omega_h} \leq \left(1 + \frac{C_a}{\alpha_0}\right) \inf_{v_h \in V_h} \|u - v_h\|_{L,\Omega_h} + \frac{1}{\alpha_0} \sup_{\phi_h \in V_h} \frac{a_h(u, \phi_h) - (f, v)_\Omega}{\|\phi_h\|_{L,\Omega_h}}.$$

(b) A unique saddle point $(u_h, \hat{\mu}_h) \in V_{\mathscr{K}} \times \hat{W}_h$ of $F_h(\cdot,\cdot)$ exists, and we have

$$\|u - u_h\|_{L,\Omega_h} \leq \tfrac{(\alpha_0 + C_a)(1+\beta_0)}{\alpha_0 \beta_0} \inf_{v_h \in V_{\mathscr{K}}} \|u - v_h\|_{L,\Omega_h} + \alpha_0^{-1} \inf_{\hat{w}_h \in \hat{W}_h} \|\hat{\mu} - \hat{w}_h\|_{L^{\mathrm{ad}},\partial\Omega},$$
$$\|\hat{\mu} - \hat{\mu}_h\|_{L^{\mathrm{ad}},\partial\Omega_h} \leq \frac{C_a}{\beta_0} \|u - u_h\|_{L,\Omega_h} + \frac{1}{\beta_0} \inf_{\hat{w}_h \in \hat{W}_h} \|\hat{\mu} - \hat{w}_h\|_{L^{\mathrm{ad}},\partial\Omega}.$$

Moreover, $u_h \in V_h$, and u_h is a minimizer of $J_h(\cdot)$.

(c) A unique saddle point $(u_h, \hat{\mu}_h, \hat{u}_h) \in V_{\mathscr{K}} \times \hat{W}_{\mathscr{K}} \times \hat{V}_h$ of $\hat{F}_h(\cdot,\cdot,\cdot)$ exists. Then, $(u_h, \hat{\mu}_h) \in \mathrm{V}_h \times \hat{W}_h$, $(u_h, \hat{\mu}_h)$ is a saddle point of of $F_h(\cdot,\cdot)$, and $\gamma_h(\hat{u}_h, w_h) = \gamma_h(u_h, w_h)$ for all $w_h \in W_h$.

The skeleton reduction. Now we define the operators $A_K \in \mathscr{L}(V_K, V_K')$, $B_K \in \mathscr{L}(V_K, \hat{W}_K')$ and $\hat{R}_K \in \mathscr{L}(\hat{V}_h, \hat{W}_K')$ and the functional $\ell_K \in V_K'$ by

$$\langle A_K v_K, \tilde{v}_K\rangle = a_K(v_K, \tilde{v}_K), \qquad v_K,\ \tilde{v}_K \in V_K,$$
$$\langle B_K v_K, \hat{w}_K\rangle = \gamma_K(v_K, \hat{w}_K), \qquad v_K \in V_K, \hat{w}_K \in \hat{W}_K,$$
$$\langle \hat{R}_K \hat{v}_h, \hat{w}_K\rangle = \gamma_K(\hat{v}_h, \hat{w}_K), \qquad \hat{v}_h \in \hat{V}_h, \hat{w}_K \in \hat{W}_K,$$
$$\langle \ell_K, v_K\rangle = (f, v_K)_K, \qquad v_K \in V_K.$$

Then, a critical point $(u_h, \hat{\mu}_h, \hat{u}_h) \in V_{\mathscr{K}} \times \hat{W}_{\mathscr{K}} \times \hat{V}_h$ of the Lagrange functional

$$\hat{F}_h(v_h, \hat{w}_h, \hat{v}_h) = \sum_K \frac{1}{2}\langle A_K v_K, v_K\rangle - \langle \ell_K, v_K\rangle + \langle B_K v_K - \hat{R}_K \hat{v}_h, \hat{w}_K\rangle$$

is characterized by

$$\begin{aligned} A_K u_K - \ell_K + B_K' \hat{\mu}_K &= 0\,,\\ B_K u_K - \hat{R}_K \hat{u}_h &= 0 \end{aligned}$$

for all K and

$$\sum_K \hat{R}_K' \hat{\mu}_K = 0\,. \tag{5}$$

This yields locally for K

$$\begin{pmatrix} A_K & B_K' \\ B_K & 0 \end{pmatrix} \begin{pmatrix} u_K \\ \hat{\mu}_K \end{pmatrix} = \begin{pmatrix} \ell_K \\ \hat{R}_K \hat{u}_h \end{pmatrix}. \tag{6}$$

Inserting the solution of (6) in (5) reduces the global saddle point problem to the self-adjoint Schur complement system $\hat{S}_h \hat{u}_h = \hat{\ell}_h$ with

$$\hat{S}_h = -\sum_K \begin{pmatrix} 0 \\ \hat{R}_K \end{pmatrix}' \begin{pmatrix} A_K & B_K' \\ B_K & 0 \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ \hat{R}_K \end{pmatrix}, \ \hat{\ell}_h = \sum_K \begin{pmatrix} 0 \\ \hat{R}_K \end{pmatrix}' \begin{pmatrix} A_K & B_K \\ B_K' & 0 \end{pmatrix}^{-1} \begin{pmatrix} \ell_K \\ 0 \end{pmatrix}.$$

Lemma 4 (similar to Lem. 6 in [7]) *The local problems* (6) *are well-posed and* $\hat{S}_h \in \mathscr{L}(\hat{V}_h, \hat{V}_h')$ *satisfies the spectral bounds*

$$\alpha_0 \|\hat{v}_h\|^2_{L,\partial\Omega_h} \le \langle \hat{S}_h \hat{v}_h, \hat{v}_h\rangle \le \frac{C_a^2}{\alpha_0^2 \beta_0^2} \|\hat{v}_h\|^2_{L,\partial\Omega_h}\,, \qquad \hat{v}_h \in \hat{V}_h\,.$$

Thus, the Schur complement system has a unique solution $\hat{u}_h$, and the local solutions $(u_K, \hat{\mu}_K)$ can be reconstructed from (6).

3 Minimal Residuals

Now we consider the DPG method in the framework of abstract trace spaces. Therefore, we introduce the minimal residual functional

$$\hat{J}(v, \hat{v}) = \sup_{w \in \mathrm{H}(L^{\mathrm{ad}}, \Omega_h)} \frac{(v, L^{\mathrm{ad}} w)_{\Omega_h} + \gamma_h(\hat{v}, w) - (f, w)_\Omega}{\|w\|_{L^{\mathrm{ad}}, \Omega_h}}\,, \ (v, \hat{v}) \in \mathrm{L}_2(\Omega, \mathbb{R}^N) \times \hat{V}.$$

The assumption (2) yields that L is injective and that the range $H = L(V)$ is closed. Furthermore, we assume $\mathrm{H}(L^{\mathrm{ad}}, \Omega_h) \subset L(V)$ and that L^{ad} is injective.

Lemma 5 *Let $(\mu, u, \hat{u}) \in \mathrm{H}(L^{\mathrm{ad}}, \Omega_h) \times \mathrm{L}_2(\Omega, \mathbb{R}^N) \times \hat{V}$ be a saddle point of*

$$F_h^{mr}(w, v, \hat{v}) = \frac{1}{2}\|w\|^2_{L^{\mathrm{ad}},\Omega_h} + (v, L^{\mathrm{ad}} w)_{\Omega_h} + \gamma_h(\hat{v}, w) - (f, w)_\Omega \,.$$

Then, $(u, \hat{u})$ is a minimizer of $\hat{J}(\cdot, \cdot)$ in $\mathrm{L}_2(\Omega, \mathbb{R}^N) \times \hat{V}$, Moreover, $u \in V$, $\hat{u} = u + \mathrm{H}_0(L, \Omega)$, $\mu = 0$ and $\hat{J}(\hat{u}, u) = 0$, and $Lu = f$.

To show that a unique saddle point $(\mu, u, \hat{u}) \in \mathrm{H}(L^{\mathrm{ad}}, \Omega_h) \times H \times \hat{V}$ of $F_h^{\mathrm{mr}}(\cdot, \cdot, \cdot)$ exists, we verify inf-sup stability with respect to the norm

$$\|(v, \hat{v})\|_{\Omega;L,\partial\Omega_h} = \Big(\|v\|^2_\Omega + \|\hat{v}\|^2_{L,\partial\Omega_h}\Big)^{1/2} \,.$$

Theorem 6 *We have for $(v, \hat{v}) \in H \times \hat{V}$*

$$\sup_{w \in \mathrm{H}(L^{\mathrm{ad}},\Omega_h)} \frac{(v, L^{\mathrm{ad}} w)_{\Omega_h} + \gamma_h(\hat{v}, w)}{\|w\|_{L^{\mathrm{ad}},\Omega_h}} \ge \frac{1}{\sqrt{4C_V^2 + 2}} \|(v, \hat{v})\|_{\Omega;L,\partial\Omega_h} \,.$$

Proof In the first step, we show that the operator $B \in \mathscr{L}(H \times \hat{V}, \mathrm{H}(L^{\mathrm{ad}}, \Omega_h)')$ defined by $\langle B(v, \hat{v}), w\rangle = (v, L^{\mathrm{ad}} w)_{\Omega_h} + \gamma_h(\hat{v}, w)$ is injective. Therefore, assume that $\langle B(v, \hat{v}), w\rangle = 0$ for all $w \in \mathrm{H}(L^{\mathrm{ad}}, \Omega_h)$. Then, $(v, L^{\mathrm{ad}} w)_{\Omega_h} = 0$ for $w \in \mathrm{C}_0^1(L^{\mathrm{ad}}, \Omega_h)$, i.e., $v \in \mathrm{H}(L, \Omega_h)$ and $Lv = 0$. This shows that $0 = \langle B(v, \hat{v}), w\rangle - (Lv, w)_{\Omega_h} = \gamma_h(\hat{v} - v, w)$ for all $w \in \mathrm{H}(L^{\mathrm{ad}}, \Omega_h)$. Testing with $w \in W$ yields $\gamma_h(v, w) = 0$, i.e., $v \in V$. Together with $Lv = 0$ and (2) we obtain $v = 0$. Then, $\gamma_h(\hat{v}, w) = 0$ for $w \in \mathrm{H}(L^{\mathrm{ad}}, \Omega_h)$ yields also $\hat{v} = 0$. This finally shows that B is injective. Thus, by duality it is sufficient to prove

$$\sup_{(v,\hat{v}) \in H \times \hat{V}} \frac{(v, L^{\mathrm{ad}} w)_{\Omega_h} + \gamma_h(\hat{v}, w)}{\|(v, \hat{v})\|_{\Omega;L,\partial\Omega_h}} \ge \frac{1}{\sqrt{4C_V^2 + 2}} \|w\|_{L^{\mathrm{ad}},\Omega_h} \,, \quad w \in \mathrm{H}(L^{\mathrm{ad}}, \Omega_h) \,.$$

For given $w \in \mathrm{H}(L^{\mathrm{ad}}, \Omega_h)$ we choose $v_w \in V$ with $Lv_w = w$, and we set $\hat{v}_w = v_w + \mathrm{H}_0(L, \Omega_h)$. Then, $\|(v_w, \hat{v}_w)\|^2_{\Omega;L,\partial\Omega_h} \le (C_V^2 + 2)\|w\|^2_{\Omega_h}$ and

$$\sup_{(v,\hat{v}) \in H \times \hat{V}} \frac{(v, L^{\mathrm{ad}} w)_{\Omega_h} + \gamma_h(\hat{v}, w)}{\|(v, \hat{v})\|_{\Omega;L,\partial\Omega_h}} \ge \frac{(v_w, L^{\mathrm{ad}} w)_{\Omega_h} + \gamma_h(\hat{v}_w, w)}{\|(v, \hat{v})\|_{\Omega;L,\partial\Omega_h}} \ge \frac{\|w\|_{\Omega_h}}{\sqrt{C_V^2 + 2}} .$$

Now, testing with $(v, \hat{v}) = (L^{\mathrm{ad}} w, 0)$ yields the assertion as in [7, Lem. 9]. □

Now, we select $H_{\mathscr{K}} \subset \mathrm{L}_2(\Omega_h, \mathbb{R}^N)$, $W_{\mathscr{K}} \subset \mathrm{H}(L^{\mathrm{ad}}, \Omega_h)$, and $\hat{V}_h \subset \hat{V}$.

Lemma 7 *Let* $(\mu_h, u_h, \hat{u}_h) \in W_{\mathscr{K}} \times H_{\mathscr{K}} \times \hat{V}_h$ *be a saddle point of* $F_h^{\mathrm{mr}}(\cdot,\cdot,\cdot)$. *Then* $(u_h, \hat{u}_h) \in H_{\mathscr{K}} \times \hat{V}_h$ *minimizes*

$$\hat{J}_h(v_h, \hat{v}_h) = \sup_{w_h \in W_{\mathscr{K}}} \frac{(v_h, L^{\mathrm{ad}} w_h)_{\Omega_h} + \gamma_h(\hat{v}_h, w_h) - (f, w_h)_\Omega}{\|w_h\|_{L^{\mathrm{ad}},\Omega_h}},$$

and we have $\|\mu_h\|_{L^{\mathrm{ad}},\Omega_h} = \hat{J}_h(u_h, \hat{u}_h)$.

The discretization is analyzed as in [6, Thm. 2.1] and [7, Thm. 6].

Theorem 8 *Assume that* $\hat{\beta}_0 > 0$ *exists such that for all* $(v_h, \hat{v}_h) \in H_{\mathscr{K}} \times \hat{V}_h$

$$\sup_{w_h \in W_{\mathscr{K}}} \frac{(v_h, L^{\mathrm{ad}} w_h)_{\Omega_h} + \gamma_h(\hat{v}_h, w_h)}{\|w_h\|_{L^{\mathrm{ad}},\Omega_h}} \geq \hat{\beta}_0 \|(v_h, \hat{v}_h)\|_{\Omega;L,\Omega_h}.$$

Then, a unique saddle point $(\mu_h, u_h, \hat{u}_h) \in W_{\mathscr{K}} \times H_{\mathscr{K}} \times \hat{V}_h$ *of* F_h^{mr} *exists with*

$$\|(u - u_h, \hat{u} - \hat{u}_h)\|_{\Omega;L,\partial\Omega_h} \leq \frac{2\sqrt{1 + C_V^2}}{\hat{\beta}_0} \inf_{(v_h,\hat{v}_h)\in H_{\mathscr{K}}\times\hat{V}_h} \|(u - u_h, \hat{v} - \hat{v}_h)\|_{\Omega;L,\partial\Omega_h}.$$

For the skeleton reduction of the minimal residual method, we define $D_K \in \mathscr{L}(W_K, W_K')$, $B_K \in \mathscr{L}(H_K, W_K')$, $R_K \in \mathscr{L}(\hat{V}_h, W_K')$, and $\ell_K \in W_K'$ by

$$\begin{aligned}
\langle D_K w_K, \tilde{w}_K\rangle &= (w_K, \tilde{w}_K)_{L^{\mathrm{ad}},K}, & w_K, \tilde{w}_K &\in W_K,\\
\langle B_K v_K, w_K\rangle &= (v_K, L^{\mathrm{ad}} w_K)_K, & v_K \in H_K, w_K &\in W_K,\\
\langle R_K \hat{v}_h, w_K\rangle &= \gamma_K(\hat{v}_h, w_K), & \hat{v}_h \in \hat{V}_h, w_K &\in W_K,\\
\langle \ell_K, w_K\rangle &= (f, w_K)_K, & w_K &\in W_K.
\end{aligned}$$

A critical point $(\mu_h, u_h, \hat{u}_h) \in W_{\mathscr{K}} \times H_{\mathscr{K}} \times \hat{V}_h$ of F_h^{mr} is characterized by

$$D_K\mu_K + B_K u_K + R_K \hat{u}_h - \ell_K = 0, \qquad B_K'\mu_K = 0, \qquad \sum_K R_K'\mu_K = 0.$$

This yields $\begin{pmatrix} 0 & B_K' \\ B_K & D_K \end{pmatrix}\begin{pmatrix} u_K \\ \mu_K \end{pmatrix} = \begin{pmatrix} 0 \\ \ell_K - R_K\hat{u}_h \end{pmatrix}$ and $\hat{S}_h^{\mathrm{mr}}\hat{u}_h = \hat{\ell}_h^{\mathrm{mr}}$ with

$$\hat{S}_h^{\mathrm{mr}} = \sum_K \begin{pmatrix} 0 \\ R_K \end{pmatrix}' \begin{pmatrix} 0 & B_K' \\ B_K & D_K \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ R_K \end{pmatrix}, \quad \hat{\ell}_h^{\mathrm{mr}} = \sum_K \begin{pmatrix} 0 \\ R_K \end{pmatrix}' \begin{pmatrix} 0 & B_K' \\ B_K & D_K \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ \ell_K \end{pmatrix}.$$

Lemma 9 (Lem. 6 in [7]) *We have*

$$\frac{\hat{\beta}_0^2}{\sqrt{4C_V^2+2}}\|\hat{v}_h\|_{\hat{V}_h} \le \langle \hat{S}^{mr}\hat{v}_h, \hat{v}_h\rangle \le \|\hat{v}_h\|^2_{L,\partial\Omega_h}\,, \qquad \hat{v}_h \in \hat{V}_h\,.$$

References

1. F. Brezzi, M. Fortin, *Mixed and Hybrid Finite Element Methods* (Springer, New York, 1991)
2. T. Bui-Thanh, L.F. Demkowicz, O. Ghattas, A unified discontinuous Petrov–Galerkin method and its analysis for Friedrich systems. SIAM J. Numer. Anal. **51**(4), 1933–1958 (2013)
3. C. Carstensen, L.F. Demkowicz, J. Gopalakrishnan, Breaking spaces and forms for the DPG method and applications including Maxwell equations, Technical report, 2015, ICES Report 15-18
4. L.F. Demkowicz, J. Gopalakrishnan, An overview of the discontinuous Petrov–Galerkin method, in *Recent Developments in Discontinuous Galerkin Finite Element Methods for Partial Differential Equations* (Springer, Springer International Publishing Switzerland, 2014), pp. 149–180
5. A. Ern, J.-L. Guermond, G. Caplain, An intrinsic criterion for the bijectivity of Hilbert operators related to Friedrich systems. Commun. Partial Differ. Equ. **32**(2), 317–341 (2007)
6. J. Gopalakrishnan, W. Qiu, An analysis of the practical DPG method. Math. Comput. **83**(286), 537–552 (2014)
7. C. Wieners, B. Wohlmuth, Robust operator estimates and the application to substructuring methods for first-order systems. ESAIM: M^{2}AN **48**, 161–175 (2014)

Iterative Coupling of Variational Space-Time Methods for Biot's System of Poroelasticity

Markus Bause and Uwe Köcher

Abstract In this work we present an iterative coupling scheme for the quasi-static Biot system of poroelasticity. For the discretization of the subproblems describing mechanical deformation and single-phase flow space-time finite element methods based on a discontinuous Galerkin approximation of the time variable are used. The spatial approximation of the flow problem is done by mixed finite element methods. The stability of the approach is illustrated by numerical experiments. The presented variational space-time framework is of higher order accuracy such that problems with high fluctuations become feasible. Moreover, it offers promising potential for the simulation of the fully dynamic Biot–Allard system coupling an elastic wave equation for solid's deformation with single-phase flow for fluid infiltration.

1 Introduction

Many physical and technical problems in mechanical, environmental and petroleum engineering involve interactions between flow and mechanical deformation in porous media. Important applications in environmental and petroleum engineering include waste disposal, hydraulic and thermal fracturing, carbon sequestration, subsurface incidence, compaction drive and oil recovery. In mechanical engineering applications arise in vibro-acoustic modeling for vehicle engineering. Promising possibilities also hold for biomechanics, medicine and earthquake engineering. In biomechanics, a poroelasticity model can be used to estimate tumor induced stress levels in the brain, and thereby assist in a clinical diagnostic setting. Poroelasticity helps also in the development of prosthetic devices. In medicine, the characterization of porous media such as trabecular bone is useful for diagnosing osteoporosis, bone disease that is manifested by the deterioration of bone microstructure. In earthquake engineering poroelastic computer simulations are used to design ways to mitigate liquefaction, the state in which the fluid pressure in a porous medium

M. Bause (✉) • U. Köcher
Helmut Schmidt University, Holstenhofweg 85, 22043 Hamburg, Germany
e-mail: bause@hsu-hh.de; koecher@hsu-hh.de

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_15

becomes greater than the forces holding the soil together, and this converting the solid-like structure into a more fluid-like structure.

Vivid examples (e.g., borehole damage by the shifting surface) indeed remind that the effect of fluid-induced deformations and fluid-solid interactions cannot be ignored. Therefore, the ability to simulate coupled mechanical deformations and flow in porous media phenomena is of particular importance from the point of view of physical realism. However, numerical modeling of such coupled processes is complex due to the structure of the model equations and continues to remain a challenging task [12]. In applications of practical interest the ratio of the characteristic intrinsical time length over the characteristic reservoir time scale is sometimes small. Then, in the singular limit of vanishing contrast coefficients the fully dynamic Biot–Allard system of poroelasticity (cf. [12]) simplifies to the quasi-static Biot system

$$-\nabla \cdot (\boldsymbol{\sigma}_0 + \boldsymbol{C} : \boldsymbol{\varepsilon}(\boldsymbol{u}) - b(p - p_0)\boldsymbol{I}) = \rho_b \boldsymbol{g}\,, \tag{1}$$

$$\partial_t \left(\frac{1}{M} p + \nabla \cdot (b\boldsymbol{u}) \right) + \nabla \cdot \boldsymbol{q} = f\,, \quad \boldsymbol{q} = -\frac{\boldsymbol{K}}{\eta} \left(\nabla p - \rho_f \boldsymbol{g} \right)\,, \tag{2}$$

$$p(0) = p_0\,, \quad \boldsymbol{u}(0) = \boldsymbol{0}\,, \tag{3}$$

to be satisfied in the domain $\Omega \subset \mathbb{R}^d$, with $d = 2$ or $d = 3$, for $t \in (0, T)$. The quasi-static feature is due to negligence of the solid's acceleration in (1). This prevents the applicability of the model (1), (2) to some classes of problems, e.g. the simulation of noise protection with tiny poroelastic layers.

In (1), (2), and (3) we denote by $\boldsymbol{u}$ the unknown displacement field, p the unknown fluid pressure, $\boldsymbol{\varepsilon}(\boldsymbol{u}) = (\nabla \boldsymbol{u} + (\nabla \boldsymbol{u})^\top)/2$ the linearized strain tensor, $\boldsymbol{C}$ the Gassmann rank-4 tensor of elasticity, $\boldsymbol{\sigma}_0$ the reference state stress tensor, b Biot's coefficient, $\rho_b = \phi\rho_f + (1-\phi)\rho_s$ the bulk density with porosity ϕ and fluid and solid phase density ρ_f and ρ_s, p_0 the reference state fluid pressure, M Biot's modulus and, finally, by $\boldsymbol{q}$ Darcy's velocity or the fluid flux. The second of the equations in (2) is the well-known Darcy law with permeability field $\boldsymbol{K}$ and fluid viscosity η. Further, $\boldsymbol{g}$ denotes gravity or, in general, some body force and f is a volumetric source. The quantities η, M, ρ_f and ρ_s are positive constants. The permeability field $\boldsymbol{K}$ is supposed to be a symmetric uniformly positive definite matrix. Furthermore, for any symmetric matrix $\boldsymbol{B}$ we assume that $(\boldsymbol{C}\boldsymbol{B}) : \boldsymbol{B} \geq a|\boldsymbol{B}|^2 + K_{\text{dr}}\text{tr}(\boldsymbol{B})^2$ is satisfied with some fixed constant $a > 0$, where K_{dr} is the drained bulk modulus [14]. Finally, we assume that ρ_b is independent of time and that $\rho_b \boldsymbol{g} = -\nabla \cdot \boldsymbol{\sigma}_0$ is satisfied. Then, for $\Omega = (0, L)^d$, $I = (0, T)$, under periodic boundary conditions for $\boldsymbol{u}$ and p with period L and for smooth L-periodic functions p_0, f and $\boldsymbol{\sigma}_0$ the Biot system (1), (2), and (3) admits an unique periodic solution [13]

$$\{\boldsymbol{u}, p\} \in C(\bar{I}; \boldsymbol{H}^1_{\text{per}}(\Omega) \cap \boldsymbol{L}^2_0(\Omega)) \times H^1(\Omega \times I) \cap C(\bar{I}; H^1_{\text{per}}(\Omega))\,.$$

Further, for $\boldsymbol{g} \in C_0^\infty(\mathbb{R}^+; \boldsymbol{L}_0^2(\Omega)), f \equiv 0$ and homogeneous initial conditions it is shown in [12] that the solution of Biot's system is smooth in time with

$$\{\boldsymbol{u}, p\} \in H^k(I; \boldsymbol{H}_{\text{per}}^1(\Omega)) \times H^k(I; H_{\text{per}}^1(\Omega)) \tag{4}$$

for all $k \in \mathbb{N}$. For existence and uniqueness results in the general situation involving boundary conditions for displacement and traction as well as for pressure and flux, prescribed on portions of the boundary, we refer to [15].

In this work we study the numerical approximation of the Biot system (1), (2), and (3) by means of an iterative splitting scheme. Recently, these approaches in coupling flow and mechanics in porous media have attracted researchers' interest; cf., e.g., [9, 12–15]. Their appreciable advantage is that by coupling the model components iteratively existing and highly developed modeling and simulation tools for each of the subproblems, including discretization, software implementation and linear solver technology, can be used fully. In recent works [2, 3, 10, 11] the authors presented space-time finite element methods for the numerical approximation of elastic wave propagation in composite material and single-phase porous media flow. Variational time discretization schemes have become of increasing importance since they offer appreciable advantages like the natural construction of higher order methods [6, 10] and the applicability of a posteriori error control by the dual weighted residual approach [1] for a simultaneous space-time adaptivity. Error analyses of these methods are also available [3, 5, 8]. In [2, 10] discontinuous and continuous finite element approximations of the time variable were combined with mixed finite element discretizations of the spatial variables and used for simulating single-phase porous media flow. In this work the described space-time discretizations are applied to the quasi-static Biot system within an iterative splitting scheme. In the numerical experiments a lowest order and a higher order time discretization of the family of discontinuous Galerkin methods are considered. To the best of our knowledge, the application of higher order schemes represents an innovation over previous works on Biot's system. Moreover, Biot's system serves as a building block for the fully dynamic Biot–Allard equations [12] for that our methods offer large potential.

2 Iterative Coupling Scheme and Space-Time Approximation of the Subproblems

In the literature [9, 13–15] four iterative coupling procedures, the *undrained split*, the *fixed stress*, the *drained split* and the *fixed strain split*, for solving the Biot system (1), (2), and (3) were proposed and studied. In [9] it was shown by a von Neumann stability analysis that the latter two methods exhibit stability problems. In [13] the stability and convergence were proved for the undrained split and the fixed stress split method by showing that these schemes define contraction maps with respect to appropriately chosen metrics. In this work we restrict ourselves to considering the *fixed stress split* iterative method that is useful in employing reservoir simulators; cf. e.g., [14].

The *fixed stress split* consists in imposing constant volumetric mean total stress $\sigma_v = \sigma_{v,0} + K_{\mathrm{dr}} \nabla \cdot \boldsymbol{u} - b(p - p_0)$ in the first half-time step resulting in

$$\begin{aligned} \left(\frac{1}{M} + \frac{b^2}{K_{\mathrm{dr}}}\right) \partial_t p^{k+1} + \nabla \cdot \boldsymbol{q}^{k+1} &= f - b \nabla \cdot \partial_t \boldsymbol{u}^k + \frac{b^2}{K_{\mathrm{dr}}} \partial_t p^k \,, \\ \boldsymbol{q}^{k+1} &= -\frac{\boldsymbol{K}}{\eta} \left(\nabla p^{k+1} - \rho_f \boldsymbol{g}\right) \end{aligned} \tag{5}$$

on $\Omega \times I$, where p^{k+1} is periodic in $\boldsymbol{x}$ with period L and $p^{k+1}(0) = p_0$. In each iteration step Eq. (5) is decoupled from the mechanical deformation subproblem and can be solved independently. In the second half-time step the effective deformation is then obtained by solving

$$-\nabla \cdot \left(\boldsymbol{\sigma}_0 + \boldsymbol{C} : \boldsymbol{\varepsilon}(\boldsymbol{u}^{k+1})\right) + b \nabla \cdot \left((p^{k+1} - p_0)\boldsymbol{I}\right) = \rho_b \boldsymbol{g} \tag{6}$$

on $\Omega \times I$, where $\boldsymbol{u}^{k+1}$ is periodic in $\boldsymbol{x}$ with period L and $\boldsymbol{u}^{k+1}(0) = \boldsymbol{0}$. The operator $\mathscr{S} : (\boldsymbol{u}^k, p^k) \mapsto (\boldsymbol{u}^{k+1}, p^{k+1})$ maps $\mathscr{D} = \{\{\boldsymbol{u}, p\} \in \mathscr{V} \times \mathscr{W} \mid \boldsymbol{u}(0) = \boldsymbol{0}\,,\ p(0) = 0\}$ into itself where without loss of generality it may be assumed that $p_0 = 0$. Here, $\mathscr{V} = \{\boldsymbol{z} \in C(\bar{I}; \boldsymbol{H}^1_{\mathrm{per}}(\Omega) \cap \boldsymbol{L}^2_0(\Omega)) \mid \partial_t \varepsilon(\boldsymbol{z}) \in \boldsymbol{L}^2(\Omega)\}$ and $\mathscr{W} = \{r \in H^1(\Omega \times I) \mid r \in C(\bar{I}; H^1_{\mathrm{per}}(\Omega))\}$. In [13] it is shown that $\mathscr{S}$ is a contraction mapping on $\mathscr{D}$ in a properly chosen metric such that $\mathscr{S}$ has an unique fixed point in $\mathscr{D}$. Boundary conditions involving displacement and traction as well as pressure and flux, prescribed on portions of the boundary,

$$p^{k+1} = p_D \text{ on } \Gamma_p \times I\,, \qquad \boldsymbol{q}^{k+1} \cdot \boldsymbol{n} = q_v \text{ on } \Gamma_q \times I\,, \tag{7}$$

$$\boldsymbol{u}^{k+1} = \boldsymbol{u}_D \text{ on } \Gamma_{\boldsymbol{u}} \times I\,, \qquad \boldsymbol{\sigma}(\boldsymbol{u}^{k+1}, p^{k+1})\boldsymbol{n} = \boldsymbol{t}_N \text{ on } \Gamma_{\boldsymbol{t}} \times I\,, \tag{8}$$

with the total stress $\boldsymbol{\sigma}(\boldsymbol{u}, p) = \boldsymbol{\sigma}_0 + \boldsymbol{C} : \boldsymbol{\varepsilon}(\boldsymbol{u}) - b(p - p_0)\boldsymbol{I}$, can be treated by the same analysis as presented in [13]. In terms of the boundary conditions (7) and (8) we define the function spaces $\boldsymbol{H}_0(\mathrm{div}; \Omega)) = \{\boldsymbol{v} \in \boldsymbol{H}(\mathrm{div}; \Omega)) \mid \boldsymbol{v}_{|\Gamma_q} \cdot \boldsymbol{n} = 0\}$ and $\mathscr{H}_0 := \{\boldsymbol{u} \in \boldsymbol{H}^1(\Omega) \mid \boldsymbol{u}_{|\Gamma_{\boldsymbol{u}}} = 0\}$.

Next, we shall describe the discretization of (5) and (6) equipped with the boundary conditions (7), (8), respectively, by space-time finite element methods. We decompose the time interval $(0, T]$ into N subintervals $I_n = (t_{n-1}, t_n]$, where $n \in \{1, \ldots, N\}$ and $0 = t_0 < t_1 < \cdots < t_{N-1} < t_N = T$ and $\tau = \max_{n=1,\ldots,N} (t_n - t_{n-1})$. Further we denote by $\mathscr{T}_h = \{K\}$ a finite element decomposition of mesh size h of the polyhedral domain $\overline{\Omega}$ into closed subsets K, quadrilaterals in two dimensions and hexahedrals in three dimensions. For the spatial discretization of (5) we use a mixed finite element approach. We choose the class of Raviart–Thomas elements for the two-dimensional case and the class of Raviart–Thomas–Nédélec elements in three space dimensions where $W_h^s \subset L^2(\Omega)$ and $\boldsymbol{V}_h^s \subset \boldsymbol{H}(\text{div};\Omega)$ denote the corresponding inf-sup stable pair of finite element spaces; cf. [2, 4, 16]. Here, s denotes the maximum polynomial order in each variable in the approximation of the scalar variable p on the reference cube $\hat{K} = [0,1]^d$, with $d = 2$ or $d = 3$. Further, we let $\boldsymbol{V}_{h,0}^s = \{\boldsymbol{v}_h \in \boldsymbol{V}_h^s \mid \boldsymbol{v}_{h|\Gamma_q} \cdot \boldsymbol{n} = 0\}$. For the spatial approximation of the displacement field $\boldsymbol{u}$ of (6) we discretize the space variables by means of a continuous Galerkin method with finite element space $\boldsymbol{H}_{h,0}^l = \{\boldsymbol{z}_h \in C(\overline{\Omega}) \mid \boldsymbol{z}_{h|K} \circ T_K \in \mathbb{Q}_l\,,\ \boldsymbol{z}_{h|\Gamma_{\boldsymbol{u}}} = \boldsymbol{0}\}$; cf. [16]. For the discretization of the time variable we use a discontinuous Galerkin approach [2, 10]. Then we define the fully discrete space-time finite element spaces as

$$
\begin{aligned}
\mathscr{W}_{\tau,h}^{r,s} &= \{w_{\tau,h} \in L^2(I;L^2(\Omega)) \mid w_{\tau,h|I_n} \in \mathscr{P}_r(I_n;W_h^s)\,,\ w_{\tau,h}(0) \in W_h^s\}\,,\\
\mathscr{V}_{\tau,h}^{r,s} &= \{\boldsymbol{v}_{\tau,h} \in L^2(I;\boldsymbol{H}_0(\text{div};\Omega)) \mid \boldsymbol{v}_{\tau,h|I_n} \in \mathscr{P}_r(I_n;\boldsymbol{V}_{h,0}^s)\,,\ \boldsymbol{v}_{\tau,h}(0) \in \boldsymbol{V}_{h,0}^s\}\,,\\
\mathscr{H}_{\tau,h}^{r,l} &= \{\boldsymbol{z}_{\tau,h} \in L^2(I;\mathscr{H}_0) \mid \boldsymbol{z}_{\tau,h|I_n} \in \mathscr{P}_r(I_n;\boldsymbol{H}_{h,0}^l)\,,\ \boldsymbol{z}_{\tau,h}(0) \in \boldsymbol{H}_{h,0}^l\}\,,
\end{aligned}
$$

where $\mathscr{P}_r(I_n;X)$ denotes the space of all polynomials in time up to degree $r \geq 0$ on I_n with values in X. We choose $l = s+1$ to equilibrate the convergence rates of the spatial discretization for the three unknown variables $\boldsymbol{u}$, p and $\boldsymbol{q}$; cf. [15, Part I, Thm. 5.2].

The space-time finite element approximation of the flow problem (5), (7) reads as follows: *Let $\boldsymbol{u}_{\tau,h}^k \in \boldsymbol{I}_h\boldsymbol{u}_d + \mathscr{H}_{\tau,h}^{r,s+1}$, $p_{\tau,h}^k \in \mathscr{W}_{\tau,h}^{r,s}$ be given and*

$$
l_p^k(w_{\tau,h}) = \left\langle f - b\nabla \cdot \partial_t \boldsymbol{u}_{\tau,h}^k + \beta\,\partial_t p_{\tau,h}^k, w_{\tau,h}\right\rangle\,, \tag{9}
$$

$$
l_{\boldsymbol{q}}^k(\boldsymbol{v}_{\tau,h}) = \langle \rho_f \boldsymbol{g}, \boldsymbol{v}_{\tau,h}\rangle - \langle p_D, \boldsymbol{v}_{\tau,h}\cdot\boldsymbol{n}\rangle_{L^2(\Gamma_p)} \tag{10}
$$

for $w_{\tau,h} \in \mathscr{W}_{\tau,h}^{r,s}$ and $\boldsymbol{v}_{\tau,h} \in \mathscr{V}_{\tau,h}^{r,s}$. Find $p_{\tau,h}^{k+1} \in \mathscr{W}_{\tau,h}^{r,s}$ and $\boldsymbol{q}_{\tau,h}^{k+1} \in \boldsymbol{\Pi}_h\boldsymbol{q}_v + \mathscr{V}_{\tau,h}^{r,s}$ such that

$$
\begin{aligned}
\sum_{n=1}^{N}\Bigg\{ &\int_{t_{n-1}}^{t_n} \langle \tilde{\beta}\,\partial_t p_{\tau,h}^{k+1}, w_{\tau,h}\rangle\,\mathrm{d}t + \int_{t_{n-1}}^{t_n} \langle \nabla \cdot q_{\tau,h}^{k+1}, w_{\tau,h}\rangle\,\mathrm{d}t\\
&+ \left\langle [\tilde{\beta}\,p_{\tau,h}^{k+1}]_{n-1}, w_{\tau,h}(t_{n-1}^+)\right\rangle\Bigg\} = \sum_{n=1}^{N}\Bigg\{\int_{t_{n-1}}^{t_n} l_p^k(w_{\tau,h})\,\mathrm{d}t\\
&- \left\langle b[\nabla\cdot\boldsymbol{u}_{\tau,h}^k]_{n-1}, w_{\tau,h}(t_{n-1}^+)\right\rangle + \left\langle \beta\,[p_{\tau,h}^k]_{n-1}, w_{\tau,h}(t_{n-1}^+)\right\rangle\Bigg\}\,,
\end{aligned} \tag{11}
$$

$$\sum_{n=1}^{N}\left\{\int_{t_{n-1}}^{t_n}\langle\tilde{\boldsymbol{K}}^{-1}\boldsymbol{q}_{\tau,h}^{k+1},\boldsymbol{v}_{\tau,h}\rangle\,\mathrm{d}t-\int_{t_{n-1}}^{t_n}\langle p_{\tau,h}^{k+1},\nabla\cdot\boldsymbol{v}_{\tau,h}\rangle\,\mathrm{d}t\right\}$$
$$=\sum_{n=1}^{N}\left\{\int_{t_{n-1}}^{t_n} l_{\boldsymbol{q}}^{k}(\boldsymbol{v}_{\tau,h})\,\mathrm{d}t\right\} \tag{12}$$

for all $w_{\tau,h}\in\mathscr{W}_{\tau,h}^{r,s}$ *and* $\boldsymbol{v}_{\tau,h}\in\mathscr{V}_{\tau,h}^{r,s}$.

In the variational problem we use the abbreviations $\beta=b^2K_{\mathrm{dr}}^{-1}$, $\tilde{\beta}:=M^{-1}+\beta$, $\tilde{\boldsymbol{K}}:=\boldsymbol{K}/\eta$. We use the notation $p_{\tau,h}(t_n^{\pm})=\lim_{t\to t_n\pm 0}p_{\tau,h}(t)$ and $[p_{\tau,h}]_n=p_{\tau,h}(t_n^{+})-p_{\tau,h}(t_n^{-})$. Further we assume that $\boldsymbol{u}_d\in\boldsymbol{H}^1(\Omega)$ satisfies $\boldsymbol{u}_d=\boldsymbol{u}_D$ on the Dirichlet boundary part $\Gamma_{\boldsymbol{u}}$ for almost every $t\in(0,T)$ (cf. Eq. (8)) and that $\boldsymbol{I}_h$ is a suitable interpolation operator for the underlying finite element space. Similarly, we let $\boldsymbol{q}_v\in\boldsymbol{H}(\mathrm{div};\Omega)$ such that $\boldsymbol{q}_v\cdot\boldsymbol{n}=q_v$ on the Neumann (flux) part Γ_q of the boundary for almost every $t\in(0,T)$ (cf. Eq. (8)) and $\boldsymbol{\Pi}_h:\boldsymbol{H}(\mathrm{div};\Omega)\mapsto\boldsymbol{V}_h^s$ denote the usual linear interpolation operator of the mixed finite element method (cf. [4]).

The space-time finite element approximation of the problem (6), (8) of the mechanical deformation reads as follows: *Let* $p_{\tau,h}^{k+1}\in\mathscr{W}_{\tau,h}^{r,s}$ *be given and*

$$l_{\boldsymbol{u}}^{k+1}(\boldsymbol{z}_{\tau,h})=\langle\rho_b\boldsymbol{g},\boldsymbol{z}_{\tau,h}\rangle+b\langle(p_{\tau,h}^{k+1}-p_0)\boldsymbol{I},\boldsymbol{\varepsilon}(\boldsymbol{z}_{\tau,h})\rangle$$
$$-\langle\boldsymbol{\sigma}_0,\boldsymbol{\varepsilon}(\boldsymbol{z}_{\tau,h})\rangle+\langle\boldsymbol{t}_N,\boldsymbol{z}_{\tau,h}\rangle_{L^2(\Gamma_t)}\,. \tag{13}$$

for $\boldsymbol{z}_{\tau,h}\in\mathscr{H}_{\tau,h}^{r,t}$. *Find* $\boldsymbol{u}_{\tau,h}^{k+1}\in\boldsymbol{I}_h\boldsymbol{u}_d+\mathscr{H}_{\tau,h}^{r,s+1}$ *such that*

$$\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}\langle\boldsymbol{C}:\boldsymbol{\varepsilon}(\boldsymbol{u}_{\tau,h}^{k+1}),\boldsymbol{\varepsilon}(\boldsymbol{z}_{\tau,h})\rangle\,\mathrm{d}t=\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}l_{\boldsymbol{u}}^{k+1}(\boldsymbol{z}_{\tau,h})\,\mathrm{d}t \tag{14}$$

for all $\boldsymbol{z}_{\tau,h}\in\mathscr{H}_{\tau,h}^{r,s+1}$.

Algebraic formulations of (9), (10), (11), (12), (13) and (14), respectively, can be obtained along the usual lines of applying finite element methods. In particular, for the time discretization we choose a basis of Lagrangian functions with support in a single subinterval I_n only and with respect to the $r+1$ Gauss quadrature points of this subinterval. For the derivation of the algebraic forms and the presentation of efficient preconditioning and solution techniques for the resulting linear systems of equations we refer to [2, 10].

3 Numerical Experiments

Now we demonstrate the application of the iteration scheme (9), (10), (11), (12), (13), (14) and study its performance properties. For this we choose the classical Terzaghi problem of consolidation of a finite layer; cf. Fig. 1. It is an uniaxial

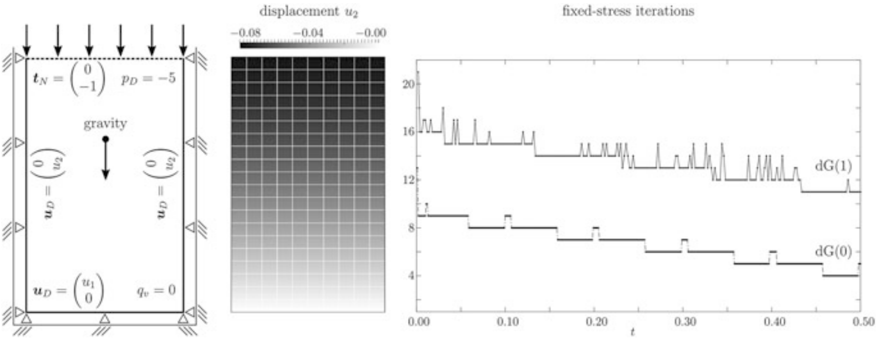


Fig. 1 Drained compaction test: problem setting (*left*), *y*-component of displacement (*center*) at $t = 0.2$ and iterations over time steps for the discontinuous time discretization with $r = 0$ (dG(0)) and $r = 1$ (dG(1)) (*right*)

compaction test in that uniform compressive traction is applied at the top surface. Exemplarily, we consider it under drained conditions where an open boundary is modeled at the top surface by prescribing a constant pressure at that portion of the boundary and zero flux conditions elsewhere. For all material parameters we use the values given in [7, Sec. 5.1]. The drained bulk modulus in (5) is chosen as $K_{\text{dr}} = 2E(1-\nu)/((1-2\nu)(1+\nu))$; cf. [14].

For our simulations we use the discontinuous Galerkin approach for the discretization of the time variable with piecewise constant and piecewise linear polynomials in time corresponding to $r = 0$ and $r = 1$, respectively, in the discrete problems (9), (10), (11), (12), (13) and (14). For $r = 0$ we choose the time step size $\tau = 1.0\text{e-}3$. For $r = 1$ we use the double step size $\tau = 2.0\text{e-}3$ in order to balance the number of unknown coefficient vectors in the algebraic systems for a macro step with $\tau = 2.0\text{e-}3$. The higher order approach is used to illustrate its feasibility and stability. In the future we plan to apply our methods to problems with a higher dynamical behaviour in time such that the higher order approach can reveal its superiority. The mixed finite element approach in (11) and (12) is applied with $s = 1$ which amounts to a piecewise bilinear approximation of the pressure variable. The temporal and spatial convergence rates thus coincide (cf. [3]) for the parameters $r = 1$ and $s = 1$ in the definition of the finite element spaces. For all iterations, i.e. for the fixed point iteration and for the iterations of the linear system solvers, we choose the tolerance `tol` $= 1.0\text{e-}10$. For the fixed point iteration this means that for each of the arising finite element coefficient vectors of the algebraic formulation (cf. [10]) the difference between two iterates, measured in the Euclidean norm, is required to be smaller than the prescribed tolerance `tol`.

In Fig. 1 we visualize the *y*-component of the displacement field $\boldsymbol{u}$ at the time $t = 0.2$. The solution agrees with the expected behaviour of the system and also with the analytical solution. The proposed iteration scheme shows a robust and stable behaviour that is documented in the right plot of Fig. 1 showing the number of fixed point iterations that are performed in each of the time steps in the interval

$t \in (0, 0.5)$ for the piecewise constant ($r = 0$; dG(0)) and piecewise linear ($r = 1$; dG(1)) approximation in time. The discontinuous Galerkin approximation in time with piecewise linear polynomials requires more fixed point iterations within the time steps, but the time step size is doubled compared with the piecewise constant approximation. Calculating for both time discretization schemes, with $r = 0$ and $r = 1$, the total number of iterations over all time steps shows that both methods perform almost the same number of fixed point iterations. Finally, we still note that the discontinuous Galerkin approach in time incorporates the initial conditions in a weak form only and thereby helps to improve the stability of the iteration in the first time step.

4 Summary

In this work we presented a space-time finite element approach to the quasi-static Biot system of poroelasticity with a discontinuous Galerkin discretization of the time variable. The approach is based on an iterative splitting scheme such that the subproblems of single-phase flow and elastic deformation are decoupled and solved sequentially. The performance properties of our approach was illustrated by a numerical experiment and for a higher order member of our family of time discretization schemes. As a work for the future we plan to compare the iterative splitting scheme with a monolithic approach in that the subproblems are solved fully coupled in a single system of equations. Moreover, we plan to use the proposed techniques for the approximation of the fully dynamic Biot–Allard system [12] for that higher order time discretization schemes offer large potential and are of significant relevance [10].

References

1. W. Bangerth, R. Rannacher, *Adaptive Methods for Differential Equations* (Birkhäuser, Basel, 2003)
2. M. Bause, U. Köcher, Variational time discretization for mixed finite element approximations of nonstationary diffusion problems. J. Comput. Appl. Math. **289**, 208–224 (2015)
3. M. Bause, F. Radu, U. Köcher, Error analysis for discretizations of parabolic problems using continuous finite elements in time and mixed finite elements in space. Numer. Math. 1–42 (2015, subm.). http://arxiv.org/abs/1504.04491
4. Z. Chen, *Finite Element Methods and Their Applications* (Springer, Berlin, 2010)
5. A. Ern, F. Schieweck, Discontinuous Galerkin method in time combined with an stabilized finite element method in space for linear first-order PDEs. Math. Comput. 1–33 (2014). Electronically published on January 11, 2016, http://dx.doi.org/10.1090/mcom/3073, http://hal.archives-ouvertes.fr/hal-00947695
6. S. Hussain, F. Schieweck, S. Turek, Higher order Galerkin time discretization for nonstationary incompressible flow, in *Numerical Mathematics and Advanced Applications 2011*, ed. by A. Cangiani et al. (Springer, Berlin, 2013), pp. 509–517

7. B. Jha, R. Juanes, A locally conservative finite element framework for the simulation of coupled flow and reservoir geomechanics. Acta Geotechnica **2**, 139–153 (2007)
8. O. Karakashin, C. Makridakis, Convergence of a continuous Galerkin method with mesh modification for nonlinear wave equations. Math. Comput. Am. Math. Soc. **74**, 85–102 (2004)
9. J. Kim, H.A. Tchelepi, R. Juanes, Stability and convergence of sequential methods for coupled flow and geomechanics: drained and undrained splits. Comput. Methods Appl. Mech. Eng. **200**, 2094–2116 (2011)
10. U. Köcher, Variational space-time methods for the elastic wave equation and the diffusion equation, PhD thesis, Helmut-Schmidt-Universität (2015). http://edoc.sub.uni-hamburg.de/hsu/volltexte/2015/3112/
11. U. Köcher, M. Bause, Variational space-time methods for the wave equation. J. Sci. Comput. **61**, 424–453 (2014)
12. A. Mikelić, M. Wheeler, Theory of the dynamic Biot–Allard equations and their link to the quasi-static Biot system. J. Math. Phys. **53**, 123702:1–15 (2012)
13. A. Mikelić, M. Wheeler, Convergence of iterative coupling for coupled flow and geomechanics. Comput. Geosci. **17**, 479–496 (2013)
14. A. Mikelić et al., Numerical convergence study of iterative coupling for coupled flow and geomechanics. Comput. Geosci. **18**, 325–341 (2014)
15. P.J. Philips, M. Wheeler, A coupling of mixed and continuous Galerkin finite element methods for poroelasticity I, II. Comput. Geosci. **11**, 131–158 (2007)
16. A. Quarteroni, A. Valli, *Numerical Approximation of Partial Differential Equations* (Springer, Berlin, 2008)

Part III
Discontinuous Galerkin Methods for PDEs

Discontinuous Galerkin Method for the Solution of Elasto-Dynamic and Fluid-Structure Interaction Problems

Miloslav Feistauer, Martin Hadrava, Adam Kosík, and Jaromír Horáček

Abstract This paper is concerned with the numerical solution of dynamic elasticity by the discontinuous Galerkin (dG) method. We consider the linear and nonlinear St. Venant-Kirchhoff model. The dynamic elasticity problem is split into two systems of first order in time. They are discretized by the discontinuous Galerkin method in space and backward difference formula in time. The developed method is tested by numerical experiments. Then the method is combined with the space-time dG method for the solution of compressible flow in a time dependent domain and used for the numerical simulation of fluid-structure interaction.

1 Description of the Dynamic Elasticity Problem

Let us consider an elastic body represented by a bounded polygonal domain $\Omega^b \subset \mathbb{R}^2$. We assume that $\partial\Omega^b = \Gamma_D^b \cup \Gamma_N^b$ and $\Gamma_D^b \cap \Gamma_N^b = \emptyset$. On Γ_D^b and Γ_N^b we prescribe the Dirichlet boundary condition and the Neumann boundary condition, respectively. The deformation of the body is described by the displacement $\boldsymbol{u} : \Omega^b \times [0, T] \rightarrow \mathbb{R}^2$ and the deformation mapping $\boldsymbol{\varphi}(\boldsymbol{X}, t) = \boldsymbol{X} + \boldsymbol{u}(\boldsymbol{X}, t)$, $\boldsymbol{X} \in \Omega^b$, $t \in [0, T]$, where $[0, T]$ with $T > 0$ is a time interval. Further, we introduce the deformation gradient $\boldsymbol{F} = \nabla\boldsymbol{\varphi}$, the Jacobian $J = \det\boldsymbol{F} > 0$ and the Green strain

M. Feistauer (✉) • M. Hadrava
Faculty of Mathematics and Physics, Charles University in Prague, Sokolovská 83, 186 75 Praha 8, Czech Republic
e-mail: feist@karlin.mff.cuni.cz; martin@hadrava.eu

A. Kosík
Faculty of Mathematics and Physics, Charles University in Prague, Sokolovská 83, 186 75 Praha 8, Czech Republic/ University of Dortmund, LS III, Vogelpothsweg 87, 44277 Dortmund, Germany
e-mail: adam.kosik.cz@gmail.com

J. Horáček
Institute of Thermomechanics, The Academy of Sciences of the Czech Republic, v. v. i., Dolejškova 1402/5, 182 00 Praha 8, Czech Republic
e-mail: jaromirh@it.cas.cz

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_16

tensor $\boldsymbol{E} \in \mathbb{R}^{2\times 2}$,

$$\boldsymbol{E} = \frac{1}{2}\left(\boldsymbol{F}^T\boldsymbol{F} - \boldsymbol{I}\right), \quad \boldsymbol{E} = (E_{ij})_{i,j=1}^2, \tag{1}$$

with the components

$$E_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial X_j} + \frac{\partial u_j}{\partial X_i}\right) + \frac{1}{2}\sum_{k=1}^{2}\frac{\partial u_k}{\partial X_i}\frac{\partial u_k}{\partial X_j}. \tag{2}$$

We set $\text{tr}(\boldsymbol{E}) = \sum_{i=1}^{2} E_{ii}$ and by $\boldsymbol{I}$ we denote the unit tensor. By the symbol Cof$\boldsymbol{F}$ we denote the cofactor of the matrix $\boldsymbol{F}$ defined as $\text{Cof}\boldsymbol{F} = J(\boldsymbol{F}^{-1})^T$. Further, we introduce the first Piola-Kirchhoff stress tensor $\boldsymbol{P}$. Its form depends on the chosen elasticity model.

The general dynamic elasticity problem is formulated in the following way: Find a displacement function $\boldsymbol{u} : \Omega^b \times [0, T] \to \mathbb{R}^2$ such that

$$\rho^b \frac{\partial^2 \boldsymbol{u}}{\partial t^2} + C_M \rho^b \frac{\partial \boldsymbol{u}}{\partial t} - \text{div}\boldsymbol{P} = \boldsymbol{f} \qquad \text{in } \Omega^b \times [0, T], \tag{3}$$

$$\boldsymbol{u} = \boldsymbol{u}_D \quad \text{in } \Gamma_D^b \times [0, T], \tag{4}$$

$$\boldsymbol{P}\boldsymbol{n} = \boldsymbol{g}_N \quad \text{in } \Gamma_N^b \times [0, T], \tag{5}$$

$$\boldsymbol{u}(\cdot, 0) = \boldsymbol{u}_0, \quad \frac{\partial \boldsymbol{u}}{\partial t}(\cdot, 0) = z_0 \qquad \text{in } \Omega^b, \tag{6}$$

where $\boldsymbol{f} : \Omega^b \times [0, T] \to \mathbb{R}^2$ is the density of the acting volume force, $\boldsymbol{g}_N : \Gamma_N^b \times [0, T] \to \mathbb{R}^2$ is the surface traction, $\boldsymbol{u}_D : \Gamma_D^b \times [0, T] \to \mathbb{R}^2$ is the prescribed displacement, $\boldsymbol{u}_0 : \Omega^b \to \mathbb{R}^2$ is the initial displacement, $z_0 : \Omega^b \to \mathbb{R}^2$ is the initial deformation velocity, $\rho^b > 0$ is the material density and $C_M \geq 0$ is the damping coefficient.

We consider two elasticity models (see [2]).

St. Venant-Kirchhoff material. In this case we set

$$\boldsymbol{\Sigma} = \lambda^b \text{tr}(\boldsymbol{E})\boldsymbol{I} + 2\mu^b \boldsymbol{E}, \quad \boldsymbol{P} = \boldsymbol{F}\boldsymbol{\Sigma}, \tag{7}$$

where $\boldsymbol{\Sigma}$ is the second Piola-Kirchhoff stress tensor. The Lamé parameters λ^b and μ^b are expressed with the aid of the Young modulus E^b and the Poisson ratio ν^b:

$$\lambda^b = \frac{E^b \nu^b}{(1 + \nu^b)(1 - 2\nu^b)}, \quad \mu^b = \frac{E^b}{2(1 + \nu^b)}. \tag{8}$$

Linear elasticity model is the simplest elasticity model obtained by the assumption of small deformations. By this assumption the second term in (2) is neglected and the linear approximation of $\boldsymbol{E}$ (linear with respect to the gradient $\boldsymbol{F}$) is denoted by $\boldsymbol{e}$ and called the small strain tensor. Then $\boldsymbol{E} = \boldsymbol{e} = (e_{ij})_{i,j=1}^2$ and

$$e_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial X_j} + \frac{\partial u_j}{\partial X_i}\right). \tag{9}$$

In this case we write

$$\boldsymbol{P} = \lambda^b \mathrm{tr}(\boldsymbol{e})\boldsymbol{I} + 2\mu^b \boldsymbol{e}. \tag{10}$$

As we see, in general, $\boldsymbol{F} = \boldsymbol{F}(\boldsymbol{u}), \boldsymbol{E} = \boldsymbol{E}(\boldsymbol{u}), \boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{u}), \boldsymbol{P} = \boldsymbol{P}(\boldsymbol{u})$.

For the time discretization of problem (3), (4), (5) and (6) we rewrite the dynamic elasticity problem as the following system of first-order in time for the displacement $\boldsymbol{u} : \Omega^b \times [0,T] \to \mathbb{R}^2$ and the deformation velocity $\boldsymbol{z} : \Omega^b \times [0,T] \to \mathbb{R}^2$:

$$\rho^b \frac{\partial \boldsymbol{z}}{\partial t} + C_M \rho^b \boldsymbol{z} - \mathrm{div}\boldsymbol{P} = \boldsymbol{f}, \quad \frac{\partial \boldsymbol{u}}{\partial t} - \boldsymbol{z} = 0 \quad \text{in } \Omega^b \times [0,T], \tag{11}$$

$$\boldsymbol{u} = \boldsymbol{u}_D \quad \text{in } \Gamma_D^b \times [0,T], \tag{12}$$

$$\boldsymbol{P}\boldsymbol{n} = \boldsymbol{g}_N \quad \text{in } \Gamma_N^b \times [0,T], \tag{13}$$

$$\boldsymbol{u}(\cdot,0) = \boldsymbol{u}_0, \quad \boldsymbol{z}(\cdot,0) = \boldsymbol{z}_0 \qquad \text{in } \Omega^b. \tag{14}$$

2 Discrete Problem

The discretization of the dynamic elasticity problem will be carried out by the dG method in space and the backward difference formula (BDF) method in time.

Let $\mathscr{T}_h^b$ be a partition of the closure $\overline{\Omega}^b$ formed by a finite number of closed triangles with disjoint interiors.

Let us consider a partition of the time interval $[0,T]$ formed by time instants $t_k = k\tau, k = 0,\dots,M$, where M is a sufficiently large positive integer and $\tau = T/M$ is the time step. (The generalization to a nonuniform partition is possible.)

Let $p > 0$ be an integer. By S_{hp} we denote the space of piecewise polynomial functions on the triangulation $\mathscr{T}_h^b$,

$$S_{hp} = \left\{v \in L^2(\Omega_h^b); v|_K \in P^p(K)\ \forall K \in \mathscr{T}_h^b\right\}, \tag{15}$$

where $P^p(K)$ denotes the space of polynomial functions of degree $\leq p$ on the element K. The approximate solution will be sought in $\boldsymbol{S}_{hp} = S_{hp} \times S_{hp}$ at each time level.

By $\mathscr{F}_h^b$ we denote the system of all faces of all elements $K \in \mathscr{T}_h^b$ and $\mathscr{F}_h^{bB}, \mathscr{F}_h^{bD}, \mathscr{F}_h^{bN}$ and $\mathscr{F}_h^{bI}$ will denote the sets of all boundary, Dirichlet, Neumann and inner faces, respectively. We set $\mathscr{F}_h^{bID} = \mathscr{F}_h^{bI} \cup \mathscr{F}_h^{bD}$. Further, for each $\Gamma \in \mathscr{F}_h^{bI}$

there exist two neighbouring elements $K_\Gamma^{(L)}, K_\Gamma^{(R)} \in \mathscr{T}_h^b$ such that $\Gamma \subset \partial K_\Gamma^{(L)} \cap \partial K_\Gamma^{(R)}$. For each $\Gamma \in \mathscr{F}_h^b$ we define a unit normal vector $\boldsymbol{n}_\Gamma$. We assume that for $\Gamma \in \mathscr{F}_h^{bB}$ the normal $\boldsymbol{n}_\Gamma$ has the same orientation as the outer normal to $\partial\Omega^b$. We use the convention that $\boldsymbol{n}_\Gamma$ is the outer normal to $\partial K_\Gamma^{(L)}$ and the inner normal to $\partial K_\Gamma^{(R)}$. For $\boldsymbol{v} \in \boldsymbol{S}_{hp}$ we introduce the following notation: $\boldsymbol{v}|_\Gamma^{(L)}$ = the trace of $\boldsymbol{v}|_{K_\Gamma^{(L)}}$ on Γ, $\boldsymbol{v}|_\Gamma^{(R)}$ = the trace of $\boldsymbol{v}|_{K_\Gamma^{(R)}}$ on Γ, $\langle \boldsymbol{v} \rangle_\Gamma = \frac{1}{2}\left(\boldsymbol{v}|_\Gamma^{(L)} + \boldsymbol{v}|_\Gamma^{(R)}\right)$, $[\boldsymbol{v}]_\Gamma = \boldsymbol{v}|_\Gamma^{(L)} - \boldsymbol{v}|_\Gamma^{(R)}$, where $\Gamma \in \mathscr{F}_h^{bI}$. If $\Gamma \in \mathscr{F}_h^{bB}$, then there exists an element $K_\Gamma^{(L)} \in \mathscr{T}_h^b$ such that $\Gamma \subset K_\Gamma^{(L)} \cap \partial\Omega_h^b$ and we set $\boldsymbol{v}|_\Gamma^{(L)}$ = the trace of $\boldsymbol{v}|_{K_\Gamma^{(L)}}$ on Γ, $\langle \boldsymbol{v} \rangle_\Gamma = [\boldsymbol{v}]_\Gamma = \boldsymbol{v}|_\Gamma^{(L)}$. Finally, we set $h_\Gamma = (h_{K_\Gamma^{(L)}} + h_{K_\Gamma^{(R)}})/2$.

In the derivation of the space discretization by the dG method the following process is essential. We multiply the governing system by a test function $\boldsymbol{v} \in \boldsymbol{S}_{hp}$, integrate the resulting relations over elements $K \in \mathscr{T}_h^b$, apply Green's theorem to the term containing $\boldsymbol{P}$, add some mutually vanishing terms, use boundary conditions and sum over all elements. In this way we get the following forms:

$$a_h^b(\boldsymbol{u}, \boldsymbol{v}) = \sum_{K \in \mathscr{T}_h^b} \int_K \boldsymbol{P}(\boldsymbol{u}) : \nabla \boldsymbol{v} \, \mathrm{d}\boldsymbol{x} - \sum_{\Gamma \in \mathscr{F}_h^{bID}} \int_\Gamma (\langle \boldsymbol{P}(\boldsymbol{u}) \rangle \boldsymbol{n}) \cdot [\boldsymbol{v}] \, \mathrm{d}S, \tag{16}$$

$$J_h^b(\boldsymbol{u}, \boldsymbol{v}) = \sum_{\Gamma \in \mathscr{F}_h^{bID}} \int_\Gamma \frac{C_W^b}{h_\Gamma} [\boldsymbol{u}] \cdot [\boldsymbol{v}] \, \mathrm{d}S, \tag{17}$$

$$\ell_h^b(\boldsymbol{v}) = \sum_{K \in \mathscr{T}_h^b} \int_K \boldsymbol{f} \cdot \boldsymbol{v} \, \mathrm{d}\boldsymbol{x} + \sum_{\Gamma \in \mathscr{F}_h^{bN}} \int_\Gamma \boldsymbol{g}_N \cdot \boldsymbol{v} \, \mathrm{d}S + \sum_{\Gamma \in \mathscr{F}_h^{bD}} \int_\Gamma \frac{C_W^b}{h_\Gamma} \boldsymbol{u}_D \cdot \boldsymbol{v} \, \mathrm{d}S, \tag{18}$$

$$A_h^b = a_h^b + J_h^b, \tag{19}$$

$$(\boldsymbol{u}, \boldsymbol{v})_{\Omega^b} = \int_{\Omega^b} \boldsymbol{u} \cdot \boldsymbol{v} \, \mathrm{d}\boldsymbol{x} = \sum_{K \in \mathscr{T}_h^b} \int_K \boldsymbol{u} \cdot \boldsymbol{v} \, \mathrm{d}\boldsymbol{x}, \tag{20}$$

where $\boldsymbol{u}, \boldsymbol{v} \in \boldsymbol{S}_{hp}$ and $C_W^b > 0$ is a sufficiently large constant.

For $k = 0, \ldots, M$ we use the approximations $\boldsymbol{u}(t_k) \approx \boldsymbol{u}_h^k \in \boldsymbol{S}_{hp}$ and $\boldsymbol{z}(t_k) \approx \boldsymbol{z}_h^k \in \boldsymbol{S}_{hp}$. A general backward difference formula approximating the time derivative reads

$$\frac{\partial \boldsymbol{u}}{\partial t}(t_{k+1}) \approx \frac{1}{\tau} \sum_{j=0}^{l} c_l \boldsymbol{u}_h^{k+1-j}, \tag{21}$$

where l is the order of the method and $c_j, j = 0, \ldots, l$, are the coefficients.

The BDF-dG approximate solution of problem (11)–(14) is defined as a couple of sequences $\{\boldsymbol{u}_h^k\}_{k=0}^M$, $\{\boldsymbol{z}_h^k\}_{k=0}^M$ such that

$$\begin{aligned}
&\text{a) } \boldsymbol{u}_h^k, \boldsymbol{z}_h^k \in \boldsymbol{S}_{hp}, \quad k = 0, \dots, M, \\
&\text{b) } \left(\frac{\rho^b}{\tau} \sum_{j=0}^{l} c_l \boldsymbol{z}_h^{k+1-j}, \boldsymbol{v}_h \right)_{\Omega_h^b} + \left(C_M \rho^b \boldsymbol{z}_h^{k+1}, \boldsymbol{v}_h \right)_{\Omega_h^b} + A_h^b(\boldsymbol{u}_h^{k+1}, \boldsymbol{v}_h) \\
&\quad = \ell_h^b(\boldsymbol{v}_h)(t_{k+1}) \quad \forall \boldsymbol{v}_h \in \boldsymbol{S}_{hp}, \\
&\text{c) } \left(\frac{\rho^b}{\tau} \sum_{j=0}^{l} c_l \boldsymbol{u}_h^{k+1-j}, \boldsymbol{v}_h \right)_{\Omega_h^b} - \left(\boldsymbol{z}_h^{k+1}, \boldsymbol{v}_h \right)_{\Omega_h^b} = 0 \quad \forall \boldsymbol{v}_h \in \boldsymbol{S}_{hp}, \\
&\quad k = 0, \dots, M-1, \\
&\text{d) } (\boldsymbol{u}_h^0 - \boldsymbol{u}_0, \boldsymbol{v}_h)_{\Omega_h^b} = 0, \quad (\boldsymbol{z}_h^0 - \boldsymbol{z}_0, \boldsymbol{v}_h)_{\Omega_h^b} = 0 \quad \forall \boldsymbol{v}_h \in \boldsymbol{S}_{hp}.
\end{aligned} \tag{22}$$

The initial values $\boldsymbol{u}_h^k$, $\boldsymbol{z}_h^k$, $k = 1, \dots, l$ are obtained by k-step BDF schemes.

In the first order BDF method we have $l = 1$, $c_0 = 1$, $c_1 = -1$ and in the second order BDF method $l = 2$ and $c_0 = 3/2$, $c_1 = -2$, $c_2 = 1/2$.

The discrete nonlinear problems are solved on each time level by the Newton method. For the solution of linear subproblems either direct UMFPACK solver or GMRES method with block diagonal preconditioning are used.

3 Numerical Experiments

3.1 A Benchmark Problem

The applicability and accuracy of the BDF-dG method is tested by the comparison with the benchmark denoted by CSM3 proposed by J. Hron and S. Turek in [4], where they used a different solution approach. We consider a 2D domain formed by the rigid cylinder with an attached elastic beam, as is shown in Fig. 1.

The following data are used: $\boldsymbol{f} = \left(0, -2\rho^b\right)^T$ [m.s^{-2}], $\rho^b = 10^3$ [kg.m^{-3}], on the left part Γ_D^b of the boundary connected with the rigid body we prescribe

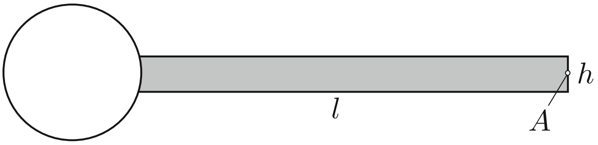


Fig. 1 Rigid cylinder with an elastic beam of the nonlinear elasticity benchmark problem

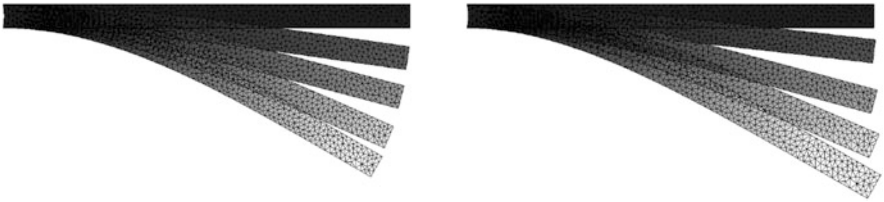


Fig. 2 The deformation of the beam in case CSM3: St. Venant-Kirchhoff model (*left*), linear elasticity model (*right*)

Table 1 CSM3: comparison of the position of the point A for BDF2, St. Venant-Kirchhoff material and different time steps τ. The values are written in the format "*mean value* ± *amplitude [frequency]*"

Method	τ	$u_1\,[\times 10^{-3}]$		$u_2\,[\times 10^{-3}]$	
Ref		-14.305 ± 14.305	[1.0995]	-63.607 ± 65.160	[1.0995]
BDF2	0.04	-10.566 ± 9.963	[1.0675]	-64.866 ± 45.218	[1.0675]
BDF2	0.02	-13.477 ± 13.462	[1.0850]	-64.133 ± 61.177	[1.0850]
BDF2	0.01	-14.119 ± 14.111	[1.0900]	-63.905 ± 64.212	[1.0900]
BDF2	0.005	-14.454 ± 14.453	[1.0925]	-64.384 ± 64.939	[1.0925]

homogeneous Dirichlet boundary condition $\boldsymbol{u}_D = \mathbf{0}$ and on the rest part Γ_N^b of the boundary we prescribe the Neumann boundary condition with no surface traction $\boldsymbol{g}_N = \mathbf{0}$. The initial conditions $\boldsymbol{u}_0 = z_0 = 0$. The material is characterized by the Young modulus $E^b = 1.4 \cdot 10^6$ and the Poisson ratio $\nu^b = 0.4$.

Figure 2 shows the deformation of the beam at several time instants computed by the linear model and St. Venant-Kirchhoff model. The linear model does not give results correct from the physical point of view in contrast to the nonlinear case. Table 1 presents the comparison between the reference results of the benchmark with our computation carried out by the second-order BDF2 time discretization with several time steps on a relatively coarse mesh with 722 elements and polynomial degree $p = 1$. According to [4], the time dependent displacement is represented by its mean value mean $= 1/2(\max + \min)$, amplitude $= 1/2(\max - \min)$ and frequency. Table 1 shows a good agreement with the reference data from [4].

3.2 *Example of Fluid-Structure Interaction*

The BDF-dG method described above is combined with the solution of compressible flow in a time dependent domain Ω_t and the resulting coupled problem is applied to the simulation of fluid-structure interaction. The boundary of Ω_t is formed by three disjoint parts: $\partial\Omega_t = \Gamma_I \cup \Gamma_O \cup \Gamma_{W_t}$, where Γ_I is the inlet, Γ_O is the outlet and Γ_{W_t} represents impermeable time-dependent walls. The time dependence of the domain

Ω_t is taken into account with the aid of the Arbitrary Lagrangian-Eulerian (ALE) method (see, e.g., [3], Chap. 10). It is based on a regular one-to-one ALE mapping of the reference configuration Ω_0 onto the current configuration Ω_t. The compressible Navier-Stokes system transformed to the ALE formulation is discretized by the space-time discontinuous Galerkin method, see [1] or [3].

In the FSI simulation the common interface between the fluid and structure at time t is defined as $\tilde{\Gamma}_{Wt} = \{\boldsymbol{x} = \boldsymbol{X} + \boldsymbol{u}(\boldsymbol{X}, t); \boldsymbol{X} \in \Gamma_N^b\} \subset \Gamma_{W_t}$. The flow and structural problems are coupled by the transmission conditions

$$\boldsymbol{v}(\boldsymbol{x}, t) = \frac{\partial \boldsymbol{u}(\boldsymbol{X}, t)}{\partial t}, \; \boldsymbol{P}(\boldsymbol{u}(\boldsymbol{X}, t))\boldsymbol{n}(\boldsymbol{x}) = \boldsymbol{\sigma}^f(\boldsymbol{x}, t)\mathrm{Cof}(\boldsymbol{F}(\boldsymbol{u}(\boldsymbol{X}, t)))\boldsymbol{n}(\boldsymbol{x}), \tag{23}$$

$$\boldsymbol{X} \in \Gamma_N^b, \; \boldsymbol{x} = \boldsymbol{X} + \boldsymbol{u}(\boldsymbol{X}, t), \quad \sigma_{ij}^f = -p\,\delta_{ij} + \lambda\delta_{ij}\mathrm{div}\boldsymbol{v} + 2\mu d_{ij}(\boldsymbol{v}),$$

$$d_{ij}(\boldsymbol{v}) = \frac{1}{2}\left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i}\right).$$

Here $\boldsymbol{v}$ denotes the fluid velocity, p is the fluid pressure and $\mu > 0$ and $\lambda = -2\mu/3$ are the fluid viscosity coefficients. The fluid-structure interaction problem is solved with the aid of a coupling procedure, see [3], Chap. 10.

As an example of the FSI problem we present the simulation of vibrations of vocal folds model excited by the airflow in a simplified geometry of vocal tract and vocal folds shown in Fig. 3. Figure 4 presents the velocity field containing a number of vortices in the deformed vocal tract at time instants $t = 0.0336, 0.0360, 0.0384, 0.0408$ s. The light shades correspond to low velocity, whereas the dark shades represent higher velocity. The pressure is in the range between 88200 and 99950 Pa. The prescribed outlet pressure is 97611 Pa. The inlet velocity is 4 ms^{-1}. The deformation of the vocal folds was computed with the use of St. Venant-Kirchhoff model. The Young modulus and the Poisson ratio have values $E^b = 12000$ Pa and $\nu^b = 0.4$, respectively, the structural damping coefficient $c_M = 1$ s^{-1} and the material density $\rho^b = 1040$ kg m^{-3}.

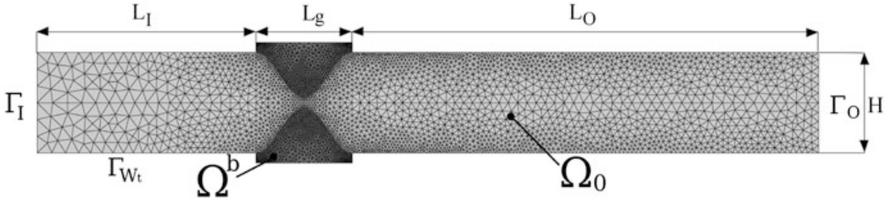


Fig. 3 Computational domain with the mesh at time $t = 0$ and the description of its size: $L_I = 50$ mm, $L_g = 15.4$ mm, $L_O = 94.6$ mm, $H = 16$ mm. The width of the channel in the narrowest part is 1.6 mm

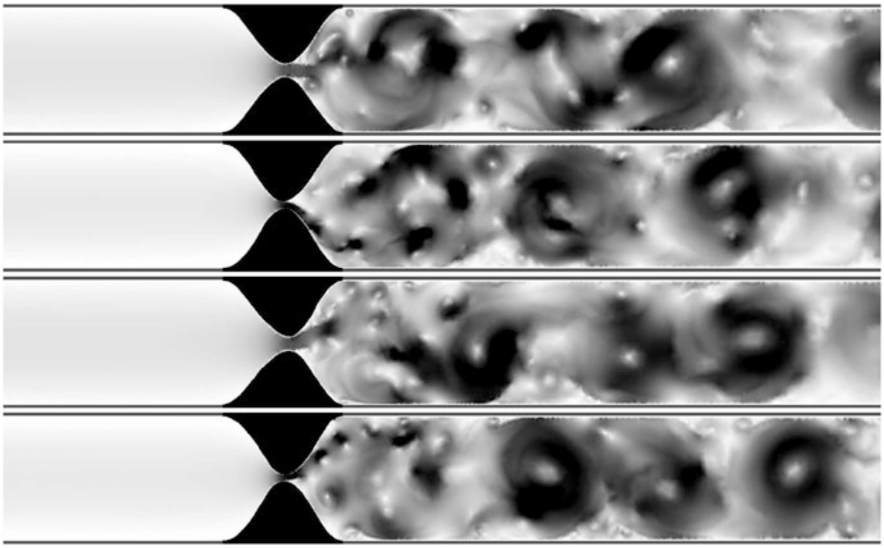


Fig. 4 The velocity field. The values of velocity magnitude (*white* to *black*) at time instans $t = 0.0336, 0.0360, 0.0384, 0.0408$ s

4 Conclusion

This paper is concerned with the application of the discontinuous Galerkin method in space combined with the BDF time discretization to the numerical solution of dynamic nonlinear elasticity problems using St. Venant-Kirchhoff material model. The method was tested on the benchmark proposed by J. Hron and S. Turek with satisfactory results. It is also shown that the method can be successfully applied to fluid-structure interaction.

Our further goal is a deeper analysis of the vocal folds vibrations using more complex geometry of vocal tract. Moreover, theoretical analysis of the developed method will be carried out.

Acknowledgements This work was supported by the grants No. 13-00522S (M. Feistauer, M. Hadrava, A. Kosík) and P101/11/0207 (J. Horáček) of the Czech Science Foundation.

References

1. J. Česenek, M. Feistauer, A. Kosík, DGFEM for the analysis of airfoil vibrations induced by compressible flow. Z. Angew. Math. Mech. **93**(6–7), 387–402 (2013)
2. P.G. Ciarlet, *Mathematical Elasticity, Volume I, Three-Dimensional Elasticity*. Volume 20 of Studies in Mathematics and Its Applications (Elsevier Science Publishers B.V., Amsterdam, 1988)

3. V. Dolejší, M. Feistauer, *Discontinuous Galerkin Method, Analysis and Applications to Compressible Flow*. Volume 48 of Springer Series in Computational Mathematics (Springer, Cham, 2015)
4. S. Turek, J. Hron, Proposal for numerical benchmarking of fluid-structure interaction between an elastic object and laminar incompressible flow, in *Fluid-Structure Interaction: Modelling, Simulation, Optimisation*, ed. by H.J. Bungartz, M. Schäfer (Springer, Berlin, 2006), pp 371–385

Stable Discontinuous Galerkin FEM Without Penalty Parameters

Lorenz John, Michael Neilan, and Iain Smears

Abstract We propose a modified local discontinuous Galerkin (LDG) method for second–order elliptic problems that does not require extrinsic penalization to ensure stability. Stability is instead achieved by showing a discrete Poincaré–Friedrichs inequality for the discrete gradient that employs a lifting of the jumps with one polynomial degree higher than the scalar approximation space. Our analysis covers rather general simplicial meshes with the possibility of hanging nodes.

1 Introduction

It is well–known that the local discontinuous Galerkin (LDG) method for second–order elliptic problems can be formulated, in part, by replacing the differential operators in the variational formulation by their discrete counterparts [3–5]. For example, on the space of discontinuous piecewise polynomials of degree at most k, the discrete gradient operator is composed of the element-wise gradient corrected by a lifting of the jumps into the space of piecewise polynomial vector fields. The original formulation of the LDG method [3] employs liftings of same polynomial degree k as the scalar finite element space, while liftings of order $k-1$ have also been considered, see the textbook [5] and the references therein. Part of the motivation for these choices of the order of the lifting is the correspondence to the order of the element-wise gradient and reasons of ease of implementation. However, unlike the continuous gradient acting on the space H_0^1, the discrete gradient operators with liftings of order $k-1$ or k fail to satisfy a discrete Poincaré–Friedrichs inequality.

L. John
Technische Universität München, 80333 München, Germany
e-mail: john@ma.tum.de

M. Neilan
University of Pittsburgh, Pittsburgh, PA 15260, USA
e-mail: neilan@pitt.edu

I. Smears (✉)
INRIA Paris-Rocquencourt, Le Chesnay, 78153, France
e-mail: iain.smears@inria.fr

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_17

Therefore, the LDG method requires additional penalization with user–defined penalty parameters to ensure stability.

In this note, we construct a modified LDG method with guaranteed stability without the need for extrinsic penalization. This result is obtained by simply increasing the polynomial degree of the lifting operator to order $k+1$ and exploiting properties of the piecewise Raviart–Thomas–Nédélec finite element space. Our analysis covers the case of meshes with hanging nodes under a mild condition of *face regularity* which we introduce in this work. We recall that the order of the lifting in the LDG method does not alter the dimension or stencil of the resulting stiffness matrix. As a result, the proposed method has a negligible increase of computational cost and inherits the advantages of the standard LDG method in terms of locality and conservativity.

The rest of the paper is organized as follows. In Sect. 2 we give the notation used throughout the manuscript and state some preliminary results. We define the lifted gradient operator with increased polynomial degree in Sect. 3 and show that the L^2 norm of this operator is equivalent to a discrete H^1 norm on piecewise polynomial spaces. We establish by means of a counterexample that the increased polynomial degree is necessary to obtain this stability estimate in Sect. 4. In Sect. 5 we propose and study the modified LDG method in the context of the Poisson equation.

2 Notation

Let $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$, be a bounded polytopal domain with Lipschitz boundary $\partial\Omega$. Let $\{\mathscr{T}_h\}_{h>0}$ be a shape- and contact–regular sequence of simplicial meshes on Ω, as defined in [5, Definition 1.38]. For each element $K \in \mathscr{T}_h$, let $h_K := \operatorname{diam} K$, with $h = \max_{K\in\mathscr{T}_h} h_K$ for each mesh $\mathscr{T}_h$. We define the faces of the mesh as in [5, Definition 1.16], and we collect all interior and boundary faces in the sets $\mathscr{F}_h^i$ and $\mathscr{F}_h^b$, respectively, and let $\mathscr{F}_h := \mathscr{F}_h^i \cup \mathscr{F}_h^b$ denote the skeleton of $\mathscr{T}_h$. In particular, $F \in \mathscr{F}_h^i$ if F has positive $(d-1)$-dimensional Hausdorff measure and if $F = \partial K_1 \cap \partial K_2$ for two distinct mesh elements K_1 and K_2. For an element $K \in \mathscr{T}_h$, we denote $\mathscr{F}(K)$ the set of faces of K, i.e. $E \in \mathscr{F}(K)$ if E is the closed convex hull of d vertices of the simplex K. Note that on a mesh with hanging nodes, a mesh face may be a proper subset of an element face, see Fig. 1, hence the notions of mesh faces and element faces do not need to coincide. In this work, the meshes are allowed to have hanging nodes, provided that they satisfy the following notion of face regularity.

Definition 1 A face $F \in \mathscr{F}_h$ is called *regular* with respect to the element K if $F \in \mathscr{F}(K)$. We say that the mesh $\mathscr{T}_h$ is *face regular* if every face of $\mathscr{F}_h$ is a regular face with respect to at least one element of $\mathscr{T}_h$.

Figure 1 illustrates the notion of face regularity with two examples. We remark that any matching mesh is face regular. On a face regular mesh, any boundary face is necessarily regular with respect to the element to which it belongs. It appears that

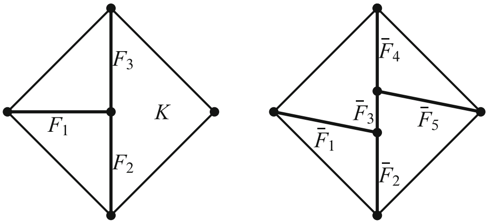


Fig. 1 Face regularity of meshes: the mesh on the left has interior faces $\mathscr{F}_h^i = \{F_i\}_{i=1}^3$, each of which is regular to at least one element in the sense of Definition 1, even though F_2 and F_3 fail to be regular with respect to the element K, since F_2 and F_3 are only proper subsets of the elemental face $F_2 \cup F_3$. Since all boundary faces are also regular, the mesh on the left is face regular in the sense of Definition 1, whereas the mesh on the right is not: the mesh face $\bar{F}_3$ fails to be regular with respect to any element of the mesh

meshes of practical interest are most likely to be face regular, so this restriction is rather mild in practice.

For integrable functions ϕ defined piecewise on either $\mathscr{T}_h$ or $\mathscr{F}_h$, we use the convention

$$\int_\Omega \phi \,\mathrm{d}x = \sum_{K\in\mathscr{T}_h} \int_K \phi \,\mathrm{d}x, \quad \int_{\mathscr{F}_h} \phi \,\mathrm{d}s = \sum_{F\in\mathscr{F}_h} \int_F \phi \,\mathrm{d}s.$$

For the integer $k \geq 1$, we define the discontinuous finite element spaces $V_{h,k}$ as the space of real-valued piecewise-polynomials of degree at most k on $\mathscr{T}_h$, and $\boldsymbol{\Sigma}_{h,k+1}$ the space of vector-valued piecewise-polynomials of degree at most $k+1$ on $\mathscr{T}_h$. We define the mesh-dependent norm $\|\cdot\|_{1,h}$ on $V_{h,k}$ by

$$\|v_h\|_{1,h}^2 := \sum_{K\in\mathscr{T}_h} \|\nabla v_h\|_{L^2(K)}^2 + \sum_{F\in\mathscr{F}_h} \frac{1}{h_F} \|[\![v_h]\!]\|_{L^2(F)}^2 \quad \forall\, v_h \in V_{h,k}, \tag{1}$$

where $h_F := \operatorname{diam} F$ for each face $F \in \mathscr{F}_h$.

We shall also make use of the (local) Raviart–Thomas–Nédélec space [7] defined by

$$\boldsymbol{RTN}_{k+1}(K) := \boldsymbol{\mathscr{P}}_k(K) \oplus \tilde{\mathscr{P}}_k(K)\,\boldsymbol{x} \subset \boldsymbol{\mathscr{P}}_{k+1}(K),$$

where $\boldsymbol{\mathscr{P}}_k(K)$ is the space of vector-valued polynomials of degree at most k on K, and $\tilde{\mathscr{P}}_k(K)$ is the space of real-valued homogeneous polynomials of degree k on K. We recall that $\boldsymbol{\tau}_h \in \boldsymbol{RTN}_{k+1}(K)$ is uniquely determined by the moments $\int_K \boldsymbol{\tau}_h \cdot \boldsymbol{\mu}_h \,\mathrm{d}x$ and $\int_E (\boldsymbol{\tau}_h \cdot \boldsymbol{n}_E)\, v_h \,\mathrm{d}s$ for all $\boldsymbol{\mu}_h \in \boldsymbol{\mathscr{P}}_{k-1}(K)$ and $v_h \in \mathscr{P}_k(E)$ for each $E \in \mathscr{F}(K)$, where $\boldsymbol{n}_E$ denotes a unit normal vector of E. We also recall that if all facial moments of $\boldsymbol{\tau}_h$ vanish on an elemental face E, then $\boldsymbol{\tau}_h \cdot \boldsymbol{n}_E$ vanishes identically on E.

For a face $F \in \mathscr{F}_h$ belonging to an element K_{ext}, we define the jump and average operators by

$$\begin{aligned} [\![w]\!]|_F &:= w|_{K_{\text{ext}}} - w|_{K_{\text{int}}}, & \{w\}|_F &:= \tfrac{1}{2}\left(w|_{K_{\text{ext}}} + w|_{K_{\text{int}}}\right), & \text{if } F \in \mathscr{F}_h^i, \\ [\![w]\!]|_F &:= w|_{K_{\text{ext}}}, & \{w\}|_F &:= w|_{K_{\text{ext}}}, & \text{if } F \in \mathscr{F}_h^b, \end{aligned}$$

where w is a sufficiently regular scalar or vector-valued function, and in the case where $F \in \mathscr{F}_h^i$, K_{int} is such that $F = \partial K_{\text{ext}} \cap \partial K_{\text{int}}$. Here, the labelling is chosen so that $\boldsymbol{n}_F$ is outward pointing with respect to K_{ext} and inward pointing with respect to K_{int}. Let $\phi \in L^2(\mathscr{F}_h)$, then the lifting operators $\boldsymbol{r}_h : L^2(\mathscr{F}_h) \to \boldsymbol{\Sigma}_{h,k+1}$ and $r_h : L^2(\mathscr{F}_h) \to V_{h,k}$ are defined by

$$\int_\Omega \boldsymbol{r}_h(\phi) \cdot \boldsymbol{\sigma}_h \,\mathrm{d}x = \int_{\mathscr{F}_h} \phi \,\{\boldsymbol{\sigma}_h \cdot \boldsymbol{n}_F\}\,\mathrm{d}s \qquad \forall\, \boldsymbol{\sigma}_h \in \boldsymbol{\Sigma}_{h,k+1}, \tag{2a}$$

$$\int_\Omega r_h(\phi)\, v_h \,\mathrm{d}x = \int_{\mathscr{F}_h^i} \phi \,\{v_h\}\,\mathrm{d}s \qquad \forall\, v_h \in V_{h,k}. \tag{2b}$$

For quantities a and b, we write $a \lesssim b$ if and only if there is a positive constant C such that $a \le Cb$, where C is independent of the quantities of interest, such as the element sizes, but possibly dependent on the shape-regularity parameters and polynomial degrees.

3 Stability of Lifted Gradients

We define the lifted gradient $G_h : V_{h,k} \to \boldsymbol{\Sigma}_{h,k+1}$ by

$$G_h(v_h) = \nabla_h v_h - \boldsymbol{r}_h([\![v_h]\!]) \quad \forall\, v_h \in V_{h,k}, \tag{3}$$

where ∇_h denotes the element-wise gradient operator. We note that G_h is usually defined with a lifting using polynomial degrees k or $k-1$, see for instance [5]. However, as we shall see, by increasing the polynomial degree of the lifting to $k+1$, we obtain the following key stability result.

Theorem 2 *Let $\{\mathscr{T}_h\}_{h>0}$ denote a shape regular, contact regular and face regular sequence of simplicial meshes on Ω. Let the norm $\|\cdot\|_{1,h}$ be defined by* (1) *and let the lifted gradient operator G_h be defined by* (3)*. Then, we have*

$$\|u_h\|_{1,h} \lesssim \|G_h(u_h)\|_{L^2(\Omega)} \lesssim \|u_h\|_{1,h} \quad \forall\, u_h \in V_{h,k}. \tag{4}$$

Proof The upper bound $\|G_h(u_h)\|_{L^2(\Omega)} \lesssim \|u_h\|_{1,h}$ is standard and we refer the reader to [5, Sec. 4.3] for a proof. To show the lower bound, consider an arbitrary $u_h \in V_{h,k}$. Since $G_h(u_h) \in \boldsymbol{\Sigma}_{h,k+1}$, we have

$$\|G_h(u_h)\|_{L^2(\Omega)} = \sup_{\boldsymbol{\tau}_h \in \boldsymbol{\Sigma}_{h,k+1} \setminus \{0\}} \frac{\int_\Omega G_h(u_h) \cdot \boldsymbol{\tau}_h \, \mathrm{d}x}{\|\boldsymbol{\tau}_h\|_{L^2(\Omega)}},$$

with the supremum being achieved by the choice $\boldsymbol{\tau}_h = G_h(u_h)$. Therefore, to show (4), it is sufficient to construct a $\boldsymbol{\tau}_h \in \boldsymbol{\Sigma}_{h,k+1}$ such that

$$\|u_h\|_{1,h}^2 \lesssim \int_\Omega G_h(u_h) \cdot \boldsymbol{\tau}_h \, \mathrm{d}x, \tag{5}$$

$$\|\boldsymbol{\tau}_h\|_{L^2(\Omega)} \lesssim \|u_h\|_{1,h}. \tag{6}$$

Let $\boldsymbol{\tau}_K \in \boldsymbol{RTN}_{k+1}(K)$ be defined by

$$\int_K \boldsymbol{\tau}_K \cdot \boldsymbol{\mu}_h \, \mathrm{d}x = \int_K \nabla u_h \cdot \boldsymbol{\mu}_h \, \mathrm{d}x \quad \forall \, \boldsymbol{\mu}_h \in \mathscr{P}_{k-1}(K), \tag{7a}$$

$$\int_E (\boldsymbol{\tau}_K \cdot \boldsymbol{n}_E) \, v_h \, \mathrm{d}s = \begin{cases} -\int_E \frac{1}{h_E} [\![u_h]\!] \, v_h \, \mathrm{d}s & \text{if } E \in \mathscr{F}_h, \\ 0 & \text{if } E \notin \mathscr{F}_h, \end{cases} \tag{7b}$$

where (7b) holds for all $v_h \in \mathscr{P}_k(E)$, for each element face $E \in \mathscr{F}(K)$. In particular, if the element face $E \in \mathscr{F}_h$, i.e. E is also a mesh face, then we require that $\boldsymbol{n}_E$ agrees with the choice of unit normal used to define the jump and average operators. If $E \notin \mathscr{F}_h$, then $\boldsymbol{\tau}_K \cdot \boldsymbol{n}_E$ vanishes identically on E, and the orientation of $\boldsymbol{n}_E$ on the left-hand side of (7b) does not matter. The global vector field $\boldsymbol{\tau}_h \in \boldsymbol{\Sigma}_{h,k+1}$ is defined element-wise by $\boldsymbol{\tau}_h|_K = \boldsymbol{\tau}_K$.

Since the mesh $\mathscr{T}_h$ is assumed to be face regular, for every $F \in \mathscr{F}_h$ there exists an element $K \in \mathscr{T}_h$ and an elemental face $E \in \mathscr{F}(K)$ such that $E = F$; then E satisfies the first condition in (7b). Therefore, the facts that $\{\boldsymbol{\tau}_h \cdot \boldsymbol{n}_F\}|_F$ and $[\![u_h]\!]|_F$ both belong to $\mathscr{P}_k(F)$ together with (7b) imply that for each $F \in \mathscr{F}_h$, one of only three situations may arise:

1. F is a boundary face and hence $F \in \mathscr{F}(K)$. In this case, we have $\{\boldsymbol{\tau}_h \cdot \boldsymbol{n}_F\}|_F = -h_F^{-1}[\![u_h]\!]|_F$.
2. F is an interior face which is regular with respect to both elements to which it belongs. In this case, we have $\{\boldsymbol{\tau}_h \cdot \boldsymbol{n}_F\}|_F = -h_F^{-1}[\![u_h]\!]|_F$.
3. F is an interior face which is regular with respect to only one of the elements to which it belongs. In this case, we have $\{\boldsymbol{\tau}_h \cdot \boldsymbol{n}_F\}|_F = -\frac{1}{2}h_F^{-1}[\![u_h]\!]|_F$, since $\boldsymbol{\tau}_h|_{K'} \cdot \boldsymbol{n}_F \equiv 0$ for the element K' with respect to which F is not regular.

Therefore, since $\boldsymbol{\tau}_h \in \boldsymbol{\Sigma}_{h,k+1}$, the definition of the lifting operator in (2a) implies that

$$\begin{aligned}\int_\Omega G_h(u_h)\cdot\boldsymbol{\tau}_h\,\mathrm{d}x &= \sum_{K\in\mathscr{T}_h}\int_K \nabla u_h\cdot\boldsymbol{\tau}_h\,\mathrm{d}x - \sum_{F\in\mathscr{F}_h}\int_F \{\boldsymbol{\tau}_h\cdot\boldsymbol{n}_F\}\,[\![u_h]\!]\,\mathrm{d}s\\ &\geq \sum_{K\in\mathscr{T}_h}\|\nabla u_h\|^2_{L^2(K)} + \frac{1}{2}\sum_{F\in\mathscr{F}_h}\frac{1}{h_F}\|[\![u_h]\!]\|^2_{L^2(F)}\\ &\geq \frac{1}{2}\|u_h\|^2_{1,h},\end{aligned}$$

where the second line follows from (7) and from the fact that $\nabla u_h|_K \in \mathscr{P}_{k-1}(K)$ for each $K \in \mathscr{T}_h$. Hence (5) is satisfied, and we now verify (6). A classical scaling argument using the Piola transformation [2, p. 59] yields

$$\begin{aligned}\|\boldsymbol{\tau}_h\|_{L^2(K)} \lesssim &\sup_{\boldsymbol{\mu}_h\in\mathscr{P}_{k-1}(K)\setminus\{0\}}\frac{\int_K \boldsymbol{\tau}_h\cdot\boldsymbol{\mu}_h\,\mathrm{d}x}{\|\boldsymbol{\mu}_h\|_{L^2(K)}}\\ &+\sum_{E\in\mathscr{F}(K)}\sup_{v_h\in\mathscr{P}_k(E)\setminus\{0\}}\frac{h_E^{1/2}\int_E(\boldsymbol{\tau}_h\cdot\boldsymbol{n}_E)v_h\,\mathrm{d}s}{\|v_h\|_{L^2(E)}}\quad\forall K\in\mathscr{T}_h.\end{aligned}$$

Therefore, it follows from (7) that, for each $K \in \mathscr{T}_h$,

$$\|\boldsymbol{\tau}_h\|^2_{L^2(K)} \lesssim \|\nabla u_h\|^2_{L^2(K)} + \sum_{F\in\mathscr{F}(K)\cap\mathscr{F}_h} h_F\|h_F^{-1}[\![u_h]\!]\|^2_{L^2(F)}. \tag{8}$$

Summing (8) over all elements therefore implies (6). □

4 Counterexample to Stability for Equal-Order Liftings

Theorem 2 shows the stability of the lifted gradient operator G_h provided that the lifting operator $\boldsymbol{r}_h$ has polynomial degree $k+1$. In this section, we verify by means of a counterexample that the stability estimate does not generally hold for lower-order liftings, including in particular the case of equal-order liftings, which are commonly used in practice; our example simplifies a similar counterexample in [1].

Example Let $\Omega = (-1,1)^2$, and consider the finite element space $V_{h,k}$ defined on a criss-cross mesh with four triangles, as depicted in Fig. 2, using piecewise linear

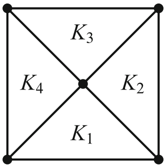


Fig. 2 Counterexample of Sect. 4: the domain $\Omega = (-1, 1)^2$ and the criss-cross mesh $\mathscr{T}_h$ considered in the example

polynomials, i.e. $k = 1$. Let $u_h \in V_{h,1}$ be the piecewise linear function defined by

$$u_h|_{K_1} = y + \frac{2}{3}, \qquad u_h|_{K_2} = x - \frac{2}{3},$$
$$u_h|_{K_3} = -y + \frac{2}{3}, \qquad u_h|_{K_4} = -x - \frac{2}{3}.$$

Direct calculations show that $\{u_h\}|_F \equiv 0$ on all interior faces $F \in \mathscr{F}_h^i$, and that $\int_K u_h \,\mathrm{d}x = 0$ for all elements $K \in \mathscr{T}_h$. Consequently, if the lifting operator $\tilde{\boldsymbol{r}}_h$ is defined in (2a) with the polynomial degree $k + 1$ replaced by k, and if $\tilde{G}_h(u_h) := \nabla_h u_h - \tilde{\boldsymbol{r}}_h([\![u_h]\!])$ denotes the equal-order lifted gradient, then we have for all $\boldsymbol{\tau}_h \in \boldsymbol{\Sigma}_{h,1}$,

$$\int_\Omega \tilde{G}_h(u_h) \cdot \boldsymbol{\tau}_h \,\mathrm{d}x = \sum_{K\in\mathscr{T}_h} \int_K \nabla_h u_h \cdot \boldsymbol{\tau}_h \,\mathrm{d}x - \sum_{F\in\mathscr{F}_h} \int_F \{\boldsymbol{\tau}_h \cdot \boldsymbol{n}_F\} [\![u_h]\!] \,\mathrm{d}s$$
$$= - \sum_{K\in\mathscr{T}_h} \int_K u_h(\nabla_h \cdot \boldsymbol{\tau}_h) \,\mathrm{d}x + \sum_{F\in\mathscr{F}_h^i} \int_F \{u_h\} [\![\boldsymbol{\tau}_h \cdot \boldsymbol{n}_F]\!] \,\mathrm{d}s = 0.$$

Since $\tilde{G}_h(u_h) \in \boldsymbol{\Sigma}_{h,1}$, we deduce that $\tilde{G}_h(u_h) = 0$, and thus it is found that no bound of the form $\|u_h\|_{1,h} \lesssim \|\tilde{G}_h(u_h)\|_{L^2(\Omega)}$ is possible. □

5 A Modified LDG Method Without Penalty Parameters

As an application of Theorem 2, consider the discretization of the homogeneous Dirichlet boundary-value problem of the Poisson equation by a modified LDG method [3, 4] as follows. For $f \in L^2(\Omega)$, let $u \in H_0^1(\Omega)$ be the unique solution of

$$\int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x = \int_\Omega f\, v \,\mathrm{d}x \quad \forall\, v \in H_0^1(\Omega). \tag{9}$$

Let the bilinear form $a_h\colon V_{h,k} \times V_{h,k} \to \mathbb{R}$ be defined by

$$a_h(u_h, v_h) = \int_\Omega G_h(u_h) \cdot G_h(v_h) \,\mathrm{d}x \quad \forall\, u_h,\, v_h \in V_{h,k}, \tag{10}$$

where the lifted gradient operator G_h was defined in (3). The bilinear form $a_h(\cdot,\cdot)$ defines a modified LDG method for (9): find $u_h \in V_{h,k}$ such that

$$a_h(u_h, v_h) = \int_\Omega f\, v_h \,\mathrm{d}x \quad \forall v_h \in V_{h,k}. \tag{11}$$

It follows from Theorem 2 that $a_h(\cdot,\cdot)$ is uniformly stable with respect to the norm $\|\cdot\|_{1,h}$, and thus (11) is well-posed for each h. Moreover, the discrete Poincaré inequality [5] implies that $\|u_h\|_{1,h} \lesssim \|f\|_{L^2(\Omega)}$ for all h, so that the numerical solutions u_h are uniformly bounded with respect to the mesh-dependent norms $\|\cdot\|_{1,h}$. The a priori error analysis for the numerical method defined by (11) may be developed following the frameworks of [3, 5, 6], although for reasons of space we do not present the arguments here.

An interesting feature of the modified LDG method (11) is that it does not require any additional stabilization, such as added penalty terms of the form $\int_{\mathscr{F}_h} \frac{\sigma_F}{h_F} [\![u_h]\!][\![v_h]\!]\,\mathrm{d}s$ for some user-defined parameter σ_F. The absence of such penalty terms enables us to show the following discrete conservation property. We define the lifted divergence $D_h \colon \boldsymbol{\Sigma}_{h,k+1} \to V_{h,k}$ by

$$D_h(\boldsymbol{\sigma}_h) = \mathrm{div}_h\, \boldsymbol{\sigma}_h - r_h([\![\boldsymbol{\sigma}_h \cdot \boldsymbol{n}_F]\!]), \quad \boldsymbol{\sigma}_h \in \boldsymbol{\Sigma}_{h,k+1}, \tag{12}$$

where div_h denotes the element-wise divergence operator, and where r_h is the scalar lifting operator defined in (2b). We note that we have the integration-by-parts identity

$$\int_\Omega \boldsymbol{\sigma}_h \cdot G_h(v_h)\,\mathrm{d}x = -\int_\Omega D_h(\boldsymbol{\sigma}_h)\, v_h \,\mathrm{d}x \quad \forall\, v_h \in V_{h,k},\ \boldsymbol{\sigma}_h \in \boldsymbol{\Sigma}_{h,k+1}, \tag{13}$$

which should be compared with the analogous continuous identity between the spaces $H^1_0(\Omega)$ and $H(\mathrm{div}, \Omega)$. Therefore, the numerical scheme (11) can be equivalently expressed in the strong form

$$-\int_\Omega D_h(G_h(u_h))\, v_h \,\mathrm{d}x = \int_\Omega f\, v_h \,\mathrm{d}x, \tag{14}$$

which implies that the numerical solution $u_h \in V_{h,k}$ solves

$$-D_h(G_h(u_h)) = \Pi^k_h f, \tag{15}$$

in the pointwise sense on each element K, where $\Pi^k_h f$ denotes the element-wise L^2-projection of f into $V_{h,k}$. Although we have shown here how the lifted gradient operator G_h of degree $k+1$ may be used to achieve a stable discretization of the Poisson equation, it is by no means restricted to this model problem, as the lifted gradients may be used to discretize the second-order terms of more general differential operators.

6 Conclusions

In this article, we studied an intrinsically stable modified LDG method without additional parameter dependent penalization. For this, we showed that increasing the degree of the lifting operator by one order leads to stability of the discrete gradient operator on face regular meshes with hanging nodes.

Acknowledgements The work of the third author was partially supported by the NSF grant DMS–1417980 and the Alfred Sloan Foundation.

References

1. F. Brezzi, M. Manzini, D. Marini, P. Pietra, A. Russo, *Discontinuous Finite Elements for Diffusion Problems*. Atti del Convegno in Memoria di F. Brioschi (Istiuto Lombardo di Scienze e Lettere, Milano, 1997)
2. D. Boffi, F. Brezzi, M. Fortin, *Mixed Finite Element Methods and Applications*. Springer Series in Computational Mathematics, vol. 44 (Springer, Berlin/New York, 2013)
3. P. Castillo, B. Cockburn, I. Perugia, D. Schötzau, An a priori error analysis of the local discontinuous Galerkin method for elliptic problems. SIAM J. Numer. Anal. **38**(5), 1676–1706 (2000) (electronic)
4. B. Cockburn, C.-W. Shu, The local discontinuous Galerkin method for time-dependent convection-diffusion systems. SIAM J. Numer. Anal. **35**(6), 2440–2463 (1998) (electronic)
5. D.A. Di Pietro, A. Ern, *Mathematical Aspects of Discontinuous Galerkin Methods*. Mathématiques & Applications, vol. 69 (Springer, Berlin/New York, 2012)
6. T. Gudi, A new error analysis for discontinuous finite element methods for linear elliptic problems. Math. Comput. **79**(272), 2169–2189 (2010)
7. J.-C. Nédélec, Mixed finite elements in $\mathbf{R}^3$. Numer. Math. **35**(3), 315–341 (1980)

Time-Space Adaptive Method of Time Layers for the Advective Allen-Cahn Equation

Murat Uzunca, Bülent Karasözen, and Ayşe Sarıaydın-Filibelioğlu

Abstract We develop an adaptive method of time layers with a linearly implicit Rosenbrock method as time integrator and symmetric interior penalty Galerkin method for space discretization for the advective Allen-Cahn equation with a non-divergence-free velocity field. Numerical simulations for advection-dominated problems demonstrate the accuracy and efficiency of the adaptive algorithm for resolving the sharp layers occurring in interface problems with small surface tension.

1 Introduction

Interfacial dynamics has great importance in modeling of multi-phase flow in material sciences. We consider the Allen-Cahn equation with advection as a model of diffuse interface for two phase flows [9]

$$\frac{\partial u}{\partial t} = \mathscr{L}u - \frac{1}{\epsilon}f(u) \quad \text{in } \Omega \times (0, T], \tag{1}$$

under homogeneous Neumann boundary condition. The elliptic linear operator $\mathscr{L}$ contains the diffusive and advective parts of the system, i.e. $\mathscr{L}u = \epsilon \Delta u - \nabla \cdot (\mathbf{V}u)$. The function $f(u) = F'(u) = 2u(1-u)(1-2u)$ stands for the cubic bistable nonlinearity with the double–well potential $F(u)$ of the two phases, and ϵ describes the surface tension. We consider a prescribed fixed velocity field $\mathbf{V} = (V_1, V_2)^T$. In coupled incompressible fluid mechanics and diffusive interface models, the velocity field satisfies the Navier–Stokes equations [9], and therefore is divergence free, i.e.

M. Uzunca (✉) • A. Sarıaydın-Filibelioğlu
Institute of Applied Mathematics, Middle East Technical University, 06800 Ankara, Turkey
e-mail: uzunca@gmail.com; uzunca@gmail.com; saayse@metu.edu.tr

B. Karasözen
Department of Mathematics & Institute of Applied Mathematics, Middle East Technical University, 06800 Ankara, Turkey
e-mail: bulent@metu.edu.tr

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_18

$\nabla \cdot \mathbf{V} = 0$. We consider in this work non-divergence-free velocity fields which are either expanding ($\nabla \cdot \mathbf{V} > 0$) or sheering ($\nabla \cdot \mathbf{V} < 0$) [9].

The dynamics of surface tension in two-phase fluids are studied numerically by different methods, among them by the level–set algorithm method and the diffuse interface method [9]. The advective Allen-Cahn equation (1) describes the diffuse interface dynamics associated with surface energies, and has two different time scales; the small surface tension, and the advective time scale. Both time scales cause computational stiffness [9].

In this work we apply the adaptive method of time layers (AMOT) [4], or adaptive Rothe method. The advective Allen-Cahn equation (1) is discretized first in time then in space, in contrast to the usual method of lines approach. The spatial discretization is considered as a perturbation of the time integration. AMOT was applied to linear and nonlinear partial differential equations using linearly implicit time integrators in several papers [3–8]. We have chosen the linearly three stage Rosenbrock (ROS3P) method [7] as the time integrator. ROS3P solver is third order convergent in time, L-stable and can efficiently deal with stiff equations. It does not show any order reduction in time in contrast to other Rosenbrock methods of order higher than two [7]. Unlike the fully implicit schemes, it requires only the solution of three linear systems per time step with the same coefficient matrix. In non-stationary models, the potential internal/boundary layers moves as the time progresses. The time step-sizes have to be adapted properly to resolve these layers accurately. The simple embedded a posteriori error estimator as the difference of second and third order ROS3P solvers allows the construction of an efficient adaptive time integrator. To resolve the sharp layers and oscillations in advection-dominated regimes, we apply the adaptive symmetric interior penalty Galerkin (SIPG) method [1, 10] for space discretization with the residual-based a posteriori error estimator [11, 12] to handle unphysical oscillations. The spatial mesh is refined or coarsened locally to obtain an accurate approximation. We show in numerical experiments that the proposed time-space algorithm AMOT is capable of damping the oscillations which may vary as the time progresses.

The paper is organized as follows. In Sect. 2 we give the fully discrete formulation of the advective AC model (1). The time-space adaptive algorithm is described in Sect. 3. In Sect. 4, results of numerical experiments for advection-dominated expanding and sheering flows are presented.

2 Time-Space Discretization

In this section we apply the method of time layers to discretize the model (1) in time. The resulting sequence of elliptic problems are discretized by the SIPG method at each time step. We consider the partition of a time interval $[0, T]$, as $I_k = (t^{k-1}, t^k]$ with time step-sizes $\tau_k = t^k - t^{k-1}$, $k = 1, 2, \ldots, J$. The approximate solution at a time $t = t^k$ is denoted by $u^k \approx u(t^k)$. We apply the 3-stage Rosenbrock solver

ROS3P [7] with an embedded error estimator in time:

$$\left(\frac{1}{\gamma\tau_k} - J^{k-1}\right) K_i = \mathscr{L}\left(z^i\right) - \frac{1}{\epsilon} f\left(z^i\right) + \sum_{j=1}^{i-1} \frac{c_{ij}}{\tau_k} K_j, \quad j = 1, 2, 3, \tag{2}$$

where $z^i = u^{k-1} + \sum_{j=1}^{i-1} a_{ij} K_j$ and $J^{k-1} := J(u^{k-1})$ is the Jacobian $J(u) = \partial_u(\mathscr{L}u - f(u)/\epsilon)$ at u^{k-1}. The second order and the third order solutions $\hat{u}^k$ and u^k, respectively, are given by

$$\begin{aligned} \hat{u}^k &= u^{k-1} + \hat{m}_1 K_1 + \hat{m}_2 K_2 + \hat{m}_3 K_3, \\ u^k &= u^{k-1} + m_1 K_1 + m_2 K_2 + m_3 K_3, \end{aligned} \tag{3}$$

with the same stage vectors K_i. The parameter values of the ROS3P solver can be found in [7]. The difference of the solutions u^k and $\hat{u}^k$ can be used as an error indicator in the time-adaptivity. Due to the linearly implicit nature of the Rosenbrock methods, the stage vectors K_i in (2) are solved by linear systems with the same coefficient matrix, which increases the computational efficiency in time integration of nonlinear PDEs [4, 6].

The semi-discrete systems (2) are discretized in space by the SIPG method with upwinding for the advective term [2]. On a time interval $I_n = (t^{k-1}, t^k]$, we consider a family $\mathscr{T}_h^k$ of shape regular elements (triangles) $E \in \mathscr{T}_h^k$. The mesh $\mathscr{T}_h^k$ is obtained by local refinement/coarsening of the mesh $\mathscr{T}_h^{k-1}$ from the previous time step. Then, with the dG finite element space $V_h^k := V_h(\mathscr{T}_h^k)$, on a time interval $I_n = (t^{k-1}, t^k]$, the fully discretized scheme of (1) reads as: for all $v_h^k \in V_h^k$, find u_h^k (or $\hat{u}_h^k$) in (3) with $K_i \in V_h^k$, $i = 1, 2, 3$, satisfying

$$\left(\left(\frac{1}{\gamma\tau_k} - J_h^{k-1}\right) K_i, v_h^k\right) = -a_h(z_h^i, v_h^k) - b_h(z_h^i, v_h^k) + \left(\sum_{j=1}^{i-1} \frac{c_{ij}}{\tau_k} K_j, v_h^k\right), \tag{4}$$

where $z_h^i = u_h^{k-1} + \sum_{j=1}^{i-1} a_{ij} K_j$, $J_h^{k-1} = J(u_h^{k-1})$ and $(\cdot, \cdot)$ stands for the discrete inner product $(\cdot, \cdot)_{L^2(\mathscr{T}_h^k)}$. The bilinear forms $a_h(u_h^k, v_h^k)$ and $b_h(u_h^k, v_h^k)$ are given by

$$\begin{aligned} a_h(u_h^k, v_h^k) = &\sum_{E \in \mathscr{T}_h^k} \int_E \epsilon \nabla u_h^k \cdot \nabla v_h^k dx + \sum_{E \in \mathscr{T}_h^k} \int_E (\mathbf{V} \cdot \nabla u_h^k + (\nabla \cdot \mathbf{V}) u_h^k) v_h^k dx \\ &+ \sum_{E \in \mathscr{T}_h^k} \int_{\partial E^- \setminus \Gamma_h^-} \mathbf{V} \cdot \mathbf{n}_E ((u_h^{out})^k - (u_h^{in})^k) v_h^k ds \\ &- \sum_{E \in \mathscr{T}_h^k} \int_{\partial E^- \cap \Gamma_h^-} \mathbf{V} \cdot \mathbf{n}_E (u_h^{in})^k v_h^k ds + \sum_{e \in \Gamma_h^k} \frac{\sigma\epsilon}{h_e} \int_e [u_h^k] \cdot [v_h^k] ds \end{aligned}$$

$$- \sum_{e\in \Gamma_h^k} \int_e (\{\epsilon \nabla v_h^k\} \cdot [u_h^k] + \{\epsilon \nabla u_h^k\} \cdot [v_h^k]) ds,$$

$$b_h(u_h^k, v_h^k) = \sum_{E\in \mathscr{T}_h^k} \int_K \frac{1}{\epsilon} f(u_h^k) v_h^k dx,$$

where u_h^{out} and u_h^{in} denote the traces on an edge from outside and inside of an element E, respectively, h_e is the length of an edge e, Γ_h^k is the set of interior edges, ∂E^- and Γ_h^- are the sets of inflow boundary edges of an element $E \in \mathscr{T}_h^k$ and on the boundary $\partial\Omega$, respectively. The parameter σ is called the penalty parameter to penalize the jumps in dG schemes, and $[\cdot]$ and $\{\cdot\}$ stand as the jump and average operators, respectively [10].

3 Adaptive Method of Time Layers (AMOT)

The AMOT scheme adjusts the time step-size and the spatial mesh adaptively on each time interval $I_k = (t^{k-1}, t^k]$. It aims, by suitable temporal and spatial estimators, to bound the total error $\|u(t^k) - \hat{u}_h^k\|$, where $u(t^k)$ is the true solution of the continuous model (1) and $\hat{u}_h^k$ is the second order (in time) discrete solution of the fully discrete system (4) on $\mathscr{T}_h^{k-1}$, at the time $t = t^k$. In order to define the temporal and spatial estimators separately, we replace the true solution $u(t^k)$ by the best available approximation $\overline{u_h^{k,+}}$ which is the third order discrete solution of the fully discrete system (4) on an auxiliary very fine mesh $\overline{\mathscr{T}_h^k} \supset \mathscr{T}_h^{k-1}$. We add and subtract in the total error the term u_h^k which is the third order (in time) discrete solution of the fully discrete system (4) on $\mathscr{T}_h^{k-1}$ at the time $t = t^k$. Then, similar to [4, Sec. 9.2], we get

$$\begin{aligned} \|u(t^k) - \hat{u}_h^k\|_{L^2(\mathscr{T}_h^{k-1})} &\approx \|\overline{u_h^{k,+}} - \hat{u}_h^k\|_{L^2(\mathscr{T}_h^{k-1})} = \|\overline{u_h^{k,+}} - u_h^k + u_h^k - \hat{u}_h^k\|_{L^2(\mathscr{T}_h^{k-1})} \\ &\leq \underbrace{\|\overline{u_h^{k,+}} - u_h^k\|_{L^2(\mathscr{T}_h^{k-1})}}_{:=\varepsilon_S} + \underbrace{\|u_h^k - \hat{u}_h^k\|_{L^2(\mathscr{T}_h^{k-1})}}_{:=\varepsilon_T} \\ &\leq TOL_S + TOL_T \leq TOL \end{aligned} \tag{5}$$

for a user prescribed tolerance TOL, and further, we set $TOL_T = \alpha TOL$ and $TOL_S = (1-\alpha)TOL$ with $0 < \alpha < 1$. In (5), the term ε_T controls the temporal adjustment, while the term ε_S controls the acceptance of spatial mesh. Note that the temporal error estimator ε_T is nothing but the difference of the second and third order solutions of the fully discrete system (4) on $\mathscr{T}_h^{k-1}$ at the time $t = t^k$. As a result, on each time interval I_k, AMOT scheme starts on the spatial mesh $\mathscr{T}_h^{k-1}$ by determining the time step-size τ_k according to the relation [4]

$$\tau^* = \sqrt[3]{\frac{\rho TOL_T}{\varepsilon_T}}\tau_k \tag{6}$$

with a safety factor $\rho \approx 0.9$, and the computed time step-size τ^* is accepted if $\varepsilon_T \leq TOL_T$.

After time step-size adjustment, AMOT scheme continues with the refinement and coarsening of the spatial mesh $\mathscr{T}_h^{k-1}$ to obtain the new spatial mesh $\mathscr{T}_h^k$ according to the spatial estimator $\varepsilon_S = \sum_E (\varepsilon_S)_E$ in (5). The local elements $E \in \mathscr{T}_h^{k-1}$ are refined for large $(\varepsilon_S)_E$ and the ones are coarsened for small $(\varepsilon_S)_E$. To determine which elements $E \in \mathscr{T}_h^{k-1}$ have to be refined, we use the condition $(\varepsilon_S)_E > 0.005 \times TOL_S$, whereas for the coarsening, we use the condition $(\varepsilon_S)_E < 10^{-13}$. For computation of the spatial estimator ε_S, we need the best available approximation $\overline{u_h^{k,+}}$ which is the solution of the discrete system (4) on a very fine auxiliary mesh $\overline{\mathscr{T}_h^k} \supset \mathscr{T}_h^{k-1}$. The auxiliary fine mesh $\overline{\mathscr{T}_h^k}$ is constructed by using a local error indicator to decide which elements $E \in \mathscr{T}_h^{k-1}$ to be refined. We use residual-based error indicator [11]

$$\eta = \left(\sum_{E \in \mathscr{T}_h^{k-1}} \eta_E^2 \right)^{1/2}, \quad \eta_E^2 = \eta_{E_R}^2 + \eta_{E_0}^2 + \eta_{E_\partial}^2, \tag{7}$$

where η_{E_R} denote the cell residuals

$$\eta_{E_R}^2 = \lambda_E^2 \left\| \frac{u_h^k - u_h^{k-1}}{\tau_k} - \epsilon \Delta u_h^k + \nabla \cdot (\mathbf{V} u_h^k) + \frac{1}{\epsilon} f(u_h^k) \right\|_{L^2(E)}^2,$$

for a weight function λ_E, while η_{E_0} and η_{E_∂} stand for the edge residuals coming from the jump of the numerical solution on the interior and Neumann boundary edges, respectively, [11, 12]. Using the local error indicators η_E in (7), we construct the auxiliary fine mesh $\overline{\mathscr{T}_h^k}$ by refining the elements $E \in M_E \subset \mathscr{T}_h^{k-1}$. To determine the set M_E, the following bulk criterion is used

$$\sum_{E \in M_E} \eta_E^2 \geq \theta \sum_{E \in \mathscr{T}_h^{k-1}} \eta_E^2$$

with a user prescribed $0 < \theta < 1$. In our simulations we take $\theta = 0.9$ since we need a very fine auxiliary mesh. The AMOT algorithm, Algorithm 1, terminates when the temporal and spatial acceptance conditions $\varepsilon_S \leq TOL_S$ and $\varepsilon_T \leq TOL_T$ are satisfied.

Algorithm 1 AMOT algorithm on a single time step $I_k = (t^{k-1}, t^k]$

```
Input: $u_h^{k-1}$, $\tau^*$, $\mathscr{T}_h^{k-1}$, $TOL_S$, $TOL_T$
Output: $u_h^{k,+}$, $\tau_k$, $\tau^*$, $\mathscr{T}_h^k$
  do
        $\tau_k = \tau^*$
        compute $u_h^k$ and $\hat{u}_h^k$ on $\mathscr{T}_h^{k-1}$
        if $\varepsilon_T > TOL_T$
            compute new step-size $\tau^*$ according to (6)
        end if
        compute error indicator $\eta$ and construct the auxiliary fine mesh $\overline{\mathscr{T}_h^k}$
        compute the best available approximation $\overline{u_h^{k,+}}$ on $\overline{\mathscr{T}_h^k}$
        if $\varepsilon_S > TOL_S$
            refine elements $E \in \mathscr{T}_h^{k-1}$ with $(\varepsilon_S)_E > 0.005 \times TOL_S$
            coarsen elements $E \in \mathscr{T}_h^{k-1}$ with $(\varepsilon_S)_E < 10^{-13}$
            construct the new spatial mesh $\mathscr{T}_h^k$
        end if
        compute $u_h^{k,+}$ on $\mathscr{T}_h^k$
  until $\varepsilon_T \leq TOL_T$ and $\varepsilon_S \leq TOL_S$
```

4 Numerical Experiments

In this section, we demonstrate the accuracy and efficiency of the proposed AMOT algorithm for expanding and sheering flow examples. In all examples, we set the tolerance $TOL = 0.001$, the parameter $\alpha = 0.5$ and the diffusion coefficient $\epsilon = 0.01$. The spatial domain is taken as $\Omega = [-1, 1]^2$ and the time interval is $[0, 06]$. For the SIPG discretization we use piecewise discontinuous linear polynomials. Numerical solutions on uniform meshes in space are computed with the constant time step-size $\tau = 0.001$ and using a 64×64 uniform spatial mesh with DoFs 24576.

4.1 Sheering Flow

We consider (1) with the sheering velocity field $\mathbf{V} = (0, -100x)$, and with the initial condition as 1 on $[-0.1, 0.1]^2$ otherwise 0 [9]. In Fig. 1, left, the unphysical oscillations of the solution on uniform mesh can be clearly seen. The oscillations are damped out by the AMOT algorithm, in Fig. 1, middle, and adaptive mesh is concentrated in the region where the sharp layers occur.

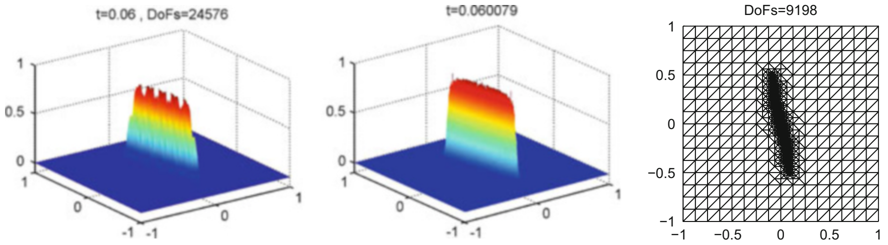


Fig. 1 Sheering flow: solution profiles at final time obtained by uniform (*left*) and adaptive (*middle*) schemes, and adaptive mesh at final time (*right*)

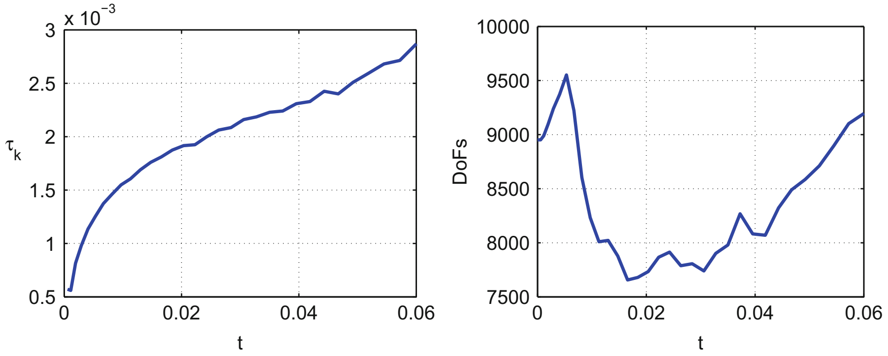


Fig. 2 Sheering flow: evolution of time step size (*left*) and DoFs (*right*)

The refinement and coarsening of AMOT algorithm works well as shown in Fig. 2, right. The mesh becomes finer at the very beginning and then, gets coarser around $t = 0.02$ as the size of the interior layer becomes smaller due to the sheering and the time step-size increases monotonically, Fig. 2, left.

4.2 Expanding Flow

As the second example, we consider the expanding velocity field $\mathbf{V} = (10x, 10y)$. The initial condition is taken as 1 in the square $[-0.3, 0.3]^2$ and 0 otherwise [9]. The unphysical oscillations are damped again in Fig. 3, middle. The mesh is refined slightly and time step-size increases at the beginning, and then refinement and coarsening proceed simultaneously, Fig. 4. Time step-size slightly decreases after $t = 0.02$ following refinement/coarsening.

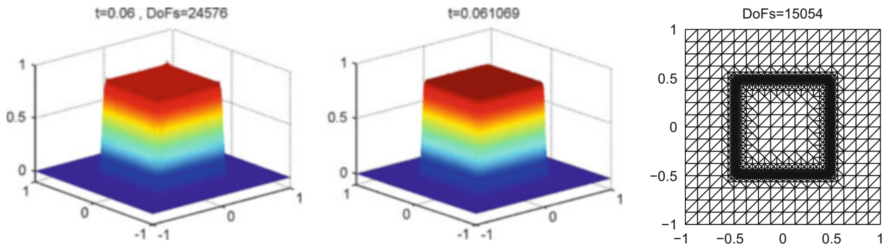


Fig. 3 Expanding flow: solution profiles at final time obtained by uniform (*left*) and adaptive (*middle*) schemes, and adaptive mesh at final time (*right*)

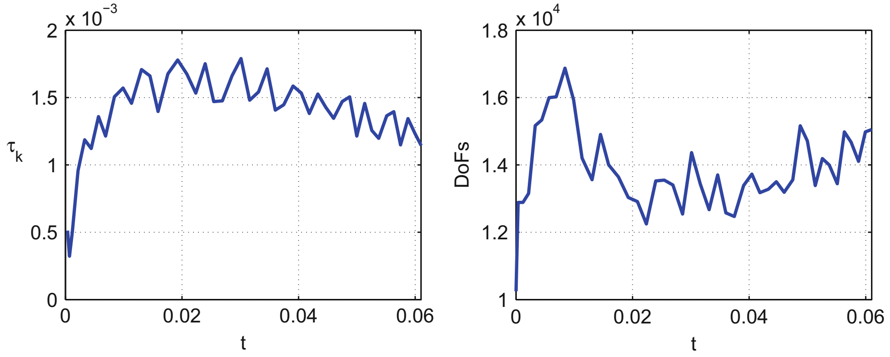


Fig. 4 Expanding flow: evolution of time step sizes (*left*) and DoFs (*right*)

Acknowledgements This work has been partially supported METU Research Fund Project BAP-07-05-2013-004 and by Scientific Human Resources Development Program (ÖYP) of the Turkish Higher Education Council (YÖK).

References

1. D. Arnold, F. Brezzi, B. Cockburn, L. Marini, Unified analysis of discontinuous Galerkin methods for elliptic problems. SIAM J. Numer. Anal. **39**, 1749–1779 (2001). doi:10.1137/S0036142901384162
2. B. Ayuso, L.D. Marini, Discontinuous Galerkin methods for advection-diffusion- reaction problems. SIAM J. Numer. Anal. **47**, 1391–1420 (2009). doi:10.1137/080719583
3. F.A. Bornemann, An adaptive multilevel approach to parabolic equations I. General theory and 1D-implementation. IMPACT Comput. Sci. Eng. **2**, 279–317 (1990)
4. P. Deuflhard, M. Weiser, *Adaptive Numerical Solution of PDEs*. De Gruyter Textbook (Walter de Gruyter, Berlin, 2012)
5. M. Frank, J. Lang, M. Schèafer, Adaptive finite element simulation of the time-dependent simplifiedP_N equations. J. Sci. Comput. **49**, 332–350 (2011). doi:10.1007/s10915-011-9466-6
6. J. Lang, *Adaptive Multilevel Solution of Nonlinear Parabolic PDE Systems. Theory, Algorithm, and Applications*. Lecture Notes in Computational Science and Engineering, vol. 16 (Springer, 2001). doi:10.1007/978-3-662-04484-1

7. J. Lang, J. Verwer, ROS3P – an accurate third-order Rosenbrock solver designed for parabolic problems. BIT Numer. Math. **41**, 731–738 (2001)
8. J. Lang, A. Walter, An adaptive Rothe method for nonlinear reaction-diffusion system. **13**, 135–146 (1993). doi:10.1016/0168-9274(93)90137-G
9. W. Liu, A. Bertozzi, T. Kolokolnikov, Diffuse interface surface tension models in an expanding flow. Commun. Math. Sci. **10**, 387–418 (2012). doi:10.4310/CMS.2012.v10.n1.a16
10. B. Rivière, *Discontinuous Galerkin Methods for Solving Elliptic and Parabolic Equations, Theory and Implementation* (SIAM, 2008). doi:10.1137/1.9780898717440
11. D. Schötzau, L. Zhu, A robust a-posteriori error estimator for discontinuous Galerkin methods for convection-diffusion equations. Appl. Numer. Math. **59**, 2236–2255 (2009). doi:10.1016/j.apnum.2008.12.014
12. M. Uzunca, B. Karasözen, M. Manguoğlu, Adaptive discontinuous Galerkin methods for non-linear diffusion-convection-reaction equations. Comput. Chem. Eng. **68**, 24–37 (2014). doi:10.1016/j.compchemeng.2014.05.002

Semi-implicit DGM Applied to a Model of Flocking

Andrea Živčáková and Václav Kučera

Abstract We present the numerical solution of a hydrodynamics model of flocking using a suitable modified semi-implicit discontinuous Galerkin method. The investigated model describing the dynamics of flocks of birds or other individual entities forming herds or swarms was introduced by Fornasier et al. (Physica D 240(1):21–31, 2011). The main idea of this model comes from the well known Cucker-Smale model. The resulting equations consist of the Euler equations for compressible flow with an additional non-local non-linear source term.

The model is discretized by the semi-implicit discontinuous Galerkin method for the compressible Euler equations of Feistauer and Kučera (J Comput Phys 224(1):208–221, 2007). We show that with a suitable treatment of the source term we can use this method for models like the model of flocking and find a numerical solution very efficiently.

1 Continuous Problem

In the paper [4], a hydrodynamic limit of a modification of the famous Cucker-Smale model is derived. The equations describe, using macroscopic quantities, the dynamics of flocks of birds or other self-organizing entities. The equations are highly nonlinear and nonlocal and are therefore extremely expensive to treat numerically, in [4] a first simple simulation was performed using the finite volume method. In this paper, we discretize the model more efficiently using the discontinuous Galerkin method.

A. Živčáková (✉) • V. Kučera
Faculty of Mathematics and Physics, Department of Numerical Mathematics, Charles University in Prague, Sokolovská 83, 186 75 Praha 8, Czech Republic
e-mail: zivcakova@karlin.mff.cuni.cz; kucera@karlin.mff.cuni.cz

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_19

Let $\Omega = (0,1) \subset \mathbb{R}$ and for $0 < L < +\infty$, we set $Q_L := \Omega \times (0, L)$. We treat the following problem: Find $\rho, u, E : Q_L \to \mathbb{R}$ such that

$$\begin{aligned} \frac{\partial \rho}{\partial t} + \operatorname{div}(\rho u) &= 0, \\ \frac{\partial \big(\rho u\big)}{\partial t} + \operatorname{div}\big(\rho u^2 + p\big) &= \lambda \mathcal{A}(\rho, u), \\ \frac{\partial E}{\partial t} + \operatorname{div}\big(u\,(E + p)\big) &= \lambda \mathcal{B}(\rho, u, T), \end{aligned} \tag{1}$$

where ρ denotes the density, u velocity, E energy, T temperature and p pressure. The right-hand side functions $\mathcal{A}$ and $\mathcal{B}$ are given by

$$\begin{aligned} \mathcal{A}(\rho, u)(x,t) &= \int_{\mathbb{R}} b(|x-y|)\Big(u(y,t) - u(x,t)\Big)\rho(x,t)\rho(y,t)\,\mathrm{d}y, \\ \mathcal{B}(\rho, u, T)(x,t) &= \int_{\mathbb{R}} b(|x-y|)\rho(x,t)\Big(\rho(y,t)u(x,t)u(y,t) - 2E(y,t)\Big)\,\mathrm{d}y, \end{aligned}$$

where

$$b(|x-y|) = \frac{K}{(\lambda + |x-y|^2)^{\beta+1}}$$

for $K, \lambda > 0$ and $\beta \geq 0$ given constants. The relations between E, p, T are

$$E = \rho\big(\tfrac{3}{2}T + \tfrac{1}{2}u^2\big), \quad p = \rho T.$$

By omitting the right-hand side terms $\mathcal{A}, \mathcal{B}$ from (1), we obtain the compressible Euler equations for a 1D monoatomic gas. In this light, we rewrite system (1) as a system of conservation laws with right-hand side source terms:

$$\frac{\partial \boldsymbol{w}}{\partial t} + \frac{\boldsymbol{f}(\boldsymbol{w})}{\partial x} = \boldsymbol{g}(\boldsymbol{w}) \ \text{ in } Q_L, \tag{2}$$

where

$$\begin{aligned} \boldsymbol{w} &= (\rho, \rho u, E)^{\top} \in \mathbb{R}^3, \\ \boldsymbol{f}(\boldsymbol{w}) &= \big(f_1(\boldsymbol{w}), f_2(\boldsymbol{w}), f_3(\boldsymbol{w})\big)^T = \big(\rho u, \rho u^2 + p, (E+p)u\big)^{\top}, \\ \boldsymbol{g}(\boldsymbol{w}) &= \lambda\big(0, \mathcal{A}(\boldsymbol{w}), \mathcal{B}(\boldsymbol{w})\big)^{\top}. \end{aligned} \tag{3}$$

The vector-valued function $\boldsymbol{w}$ is called the *state vector* and the function $\boldsymbol{f}$ is the so-called *Euler* or *inviscid flux*. In (3), we write the right-hand side terms $\mathcal{A}, \mathcal{B}$ as functions of the state vector $\boldsymbol{w}$, although in (1), they are written in terms of the nonconservative variables. Expressing $\mathcal{A}, \mathcal{B}$ in $\boldsymbol{w}$ in a suitable way is a key ingredient in our scheme and will be described in Sect. 2.3.1.

The resulting system is equipped with the initial condition

$$\boldsymbol{w}(x,0) = \boldsymbol{w}^0(x), \quad x \in \Omega,$$

and periodic boundary conditions, for simplicity.

The Euler flux is a homogeneous function, which implies

$$\boldsymbol{f}(\boldsymbol{w}) = \mathbb{A}\,(\boldsymbol{w})\boldsymbol{w}, \tag{4}$$

where $\mathbb{A} = \frac{D\boldsymbol{f}}{D\boldsymbol{w}}$. Furthermore, the Jacobi matrix of the Euler flux is *diagonally hyperbolic*. In 1D this means the matrix

$$\mathbb{P}(\boldsymbol{w}, n) := \mathbb{A}\,(\boldsymbol{w})n$$

is diagonalizable with real eigenvalues, where $n = \pm 1$. I.e. there exists a matrix $\mathbb{T}(\boldsymbol{w}, n) \in \mathbb{R}^{3,3}$ and a diagonal matrix $\mathbb{D}\,(\boldsymbol{w}, n) \in \mathbb{R}^{3,3}$ with eigenvalues $\lambda_1, \lambda_2, \lambda_3 \in \mathbb{R}$ such that

$$\mathbb{P}(\boldsymbol{w}, n) = \mathbb{T}\mathbb{D}\,\mathbb{T}^{-1}, \;\; \text{where}\; \mathbb{D}\,(\boldsymbol{w}, n) = diag(\lambda_1, \lambda_2, \lambda_3). \tag{5}$$

2 Discretization

We shall use the multidimensional notation for $\Omega \subset \mathbb{R}^d$, although in our computations we have $d = 1$. Let $\mathcal{T}_h$ be triangulation of Ω and $\mathcal{F}_h$ the system of all faces (nodes in 1D) of $\mathcal{T}_h$. For each $\Gamma \in \mathcal{F}_h$ we choose a unit normal $n_\Gamma = \pm 1$, which, for $\Gamma \subset \partial\Omega$, has the same orientation as the outer normal to Ω. For each *interior* face $\Gamma \in \mathcal{F}_h$ there exist two neighbours $K_\Gamma^{(L)}, K_\Gamma^{(R)} \in \mathcal{T}_h$ such that n_Γ is the outer normal to $K_\Gamma^{(L)}$. For v piecewise defined on $\mathcal{T}_h$ and $\Gamma \in \mathcal{F}_h$ we introduce $v|_\Gamma^{(L)}$ is the trace of $v|_{K_\Gamma^{(L)}}$ on Γ, $v|_\Gamma^{(R)}$ is the trace of $v|_{K_\Gamma^{(R)}}$ on Γ and $[v]_\Gamma = v|_\Gamma^{(L)} - v|_\Gamma^{(R)}$ is the *jump* of v. On $\partial\Omega$, we define $v|_\Gamma^{(L)}, v|_\Gamma^{(R)}$ using periodic boundary conditions. If $[\cdot]_\Gamma, v|_\Gamma^{(L)}, v|_\Gamma^{(R)}$ appear in an integral over $\Gamma \in \mathcal{F}_h$, we omit the subscript Γ.

Let $p \in \mathbb{N}$ and let $P^p(K)$ be the space of polynomials on $K \in \mathcal{T}_h$ of degree $\leq p$. The approximate solution will be sought in the space of discontinuous piecewise polynomial functions

$$\boldsymbol{S}_h := [S_h]^3, \;\; \text{where}\; S_h = \{v;\; v|_K \in P^p(K), \forall K \in \mathcal{T}_h\}.$$

2.1 Discontinuous Galerkin Space Semidiscretization

To derive the discrete problem, we assume that $\boldsymbol{w}$ is a classical solution of problem (2). We multiply (2) by a test function $\boldsymbol{\varphi}_h \in \boldsymbol{S}_h$, integrate over $K \in \mathcal{T}_h$ and apply Green's theorem in the convective terms. Summing over all $K \in \mathcal{T}_h$ and rearranging, we obtain

$$\int_\Omega \frac{\partial \boldsymbol{w}}{\partial t} \cdot \boldsymbol{\varphi}\,\mathrm{d}x + \int_{\mathcal{F}_h} \boldsymbol{f}(\boldsymbol{w})\mathbf{n} \cdot [\boldsymbol{\varphi}]\,\mathrm{d}\mathbf{S} - \sum_{\mathbf{K}\in\mathcal{T}_\mathbf{h}} \int_\mathbf{K} \boldsymbol{f}(\boldsymbol{w}) \cdot \frac{\partial \boldsymbol{\varphi}}{\partial \mathbf{x}}\,\mathrm{d}\mathbf{x} = \int_{\boldsymbol{\Omega}} \boldsymbol{g}(\boldsymbol{w}) \cdot \boldsymbol{\varphi}\,\mathrm{d}\mathbf{x}.$$

The discrete approximation of $\boldsymbol{w}$ will be sought in $\boldsymbol{S}_h$, we need to give proper meaning to the boundary integral term. Similarly as in the finite volume method, we to approximate the physical flux $\boldsymbol{f}(\boldsymbol{w})\mathbf{n}$ through an edge $\Gamma \in \mathcal{F}_h$ by a so-called *numerical flux* $\mathbf{H}(\boldsymbol{w}^{(L)}, \boldsymbol{w}^{(R)}, n)$

$$\int_{\mathcal{F}_h} \boldsymbol{f}(\boldsymbol{w})\mathbf{n} \cdot [\boldsymbol{\varphi}]\,\mathrm{d}\mathbf{S} \approx \int_{\mathcal{F}_\mathbf{h}} \mathbf{H}(\boldsymbol{w}^{(\mathbf{L})}, \boldsymbol{w}^{(\mathbf{R})}, \mathbf{n}) \cdot [\boldsymbol{\varphi}]\,\mathrm{d}\mathbf{S}. \tag{6}$$

The specific choice of $\mathbf{H}$ will be discussed in Sect. 2.2.

For $\boldsymbol{w}, \boldsymbol{\varphi} \in H^1(\Omega, \mathcal{T}_h)$, we can define the following forms.
Convective form:

$$b_h(\boldsymbol{w}, \boldsymbol{\varphi}) = \int_{\mathcal{F}_h} \mathbf{H}(\boldsymbol{w}^{(L)}, \boldsymbol{w}^{(R)}, n) \cdot [\boldsymbol{\varphi}]\,\mathrm{d}S - \sum_{K\in\mathcal{T}_h} \int_K \boldsymbol{f}(\boldsymbol{w}) \cdot \frac{\partial \boldsymbol{\varphi}}{\partial \mathbf{x}}\,\mathrm{d}\mathbf{x},$$

right-hand side source term form:

$$l_h(\boldsymbol{w}, \boldsymbol{\varphi}) = -\int_\Omega \boldsymbol{g}(\boldsymbol{w}) \cdot \boldsymbol{\varphi}\,\mathrm{d}x.$$

Finally, we introduce the space semi-discrete problem: We seek $\boldsymbol{w}_h \in C^1([0,T]; \boldsymbol{S}_h)$ such that for all $\boldsymbol{\varphi}_h \in \boldsymbol{S}_h$ and for all $t \in (0,T)$

$$\frac{d}{dt}(\boldsymbol{w}_h(t), \boldsymbol{\varphi}_h) + b_h(\boldsymbol{w}_h(t), \boldsymbol{\varphi}_h) + l_h(\boldsymbol{w}_h(t), \boldsymbol{\varphi}_h) = 0. \tag{7}$$

2.2 Numerical Flux

The choice of the numerical flux is a very important question in the finite volume and DG schemes. As such, it has been extensively studied from theoretical and practical points of view and many different constructions exist. Here, we will use the *Vijayasundaram* numerical flux, cf. [5], which is suitable for our semi-implicit time discretization. This numerical flux is based on the flux vector splitting concept, and can be viewed as an extension of the upwind numerical flux to nonlinear systems of equations. We use the diagonal hyperbolicity (5) and define the *positive* and *negative* parts of matrix $\mathbb{P}$:

$$\mathbb{P}^{\pm}(\boldsymbol{w},n) = \mathbb{T}(\boldsymbol{w},n)\mathbb{D}^{\pm}(\boldsymbol{w},n)\mathbb{T}^{-1}(\boldsymbol{w},n), \quad \mathbb{D}^{\pm}(\boldsymbol{w},n) = diag(\lambda_1^{\pm},\lambda_2^{\pm},\lambda_3^{\pm}), \tag{8}$$

where $\lambda^{+} = \max\{0,\lambda\}$, $\lambda^{-} = \min\{0,\lambda\}$. Then $\mathbb{P}(\boldsymbol{w},n) = \mathbb{P}^{+}(\boldsymbol{w},n) + \mathbb{P}^{-}(\boldsymbol{w},n)$ and we can define the Vijayasundaram numerical flux as

$$\mathbf{H}_{VS}(\boldsymbol{w}_L,\boldsymbol{w}_R,n) = \mathbb{P}^{+}\left(\tfrac{\boldsymbol{w}_L+\boldsymbol{w}_R}{2},n\right)\boldsymbol{w}_L + \mathbb{P}^{-}\left(\tfrac{\boldsymbol{w}_L+\boldsymbol{w}_R}{2},n\right)\boldsymbol{w}_R. \tag{9}$$

Explicit formulas for $\mathbb{P}, \mathbb{T}, \mathbb{T}^{-1}$ and $\mathbb{D}$ can be found e.g. in [3].

2.3 Time Discretization

After choosing some basis of the space $\boldsymbol{S}_h$, Eq. (7) represents a system of nonlinear ordinary differential equations, which must be discretized with respect to time. Due to severe time step restrictions, we want to avoid using an explicit scheme. However an implicit time discretization is also very expensive due to its nonlinearity. Therefore we choose the semi-implicit scheme of [2] as a basis and apply it to our problem.

Let $0 = t_0 < t_1 < t_2 < \ldots$ be a partition of time interval $[0,T]$ and define $\tau_k = t_{k+1} - t_k$. We approximate $\boldsymbol{w}_h^k \approx \boldsymbol{w}_h(t_k)$, where $\boldsymbol{w}_h^k \in \boldsymbol{S}_h$. We use a first order backward difference approximation for the time derivative. The resulting scheme reads

$$\left(\frac{\boldsymbol{w}_h^{k+1} - \boldsymbol{w}_h^k}{\tau_k},\boldsymbol{\varphi}_h\right) + b_h(\boldsymbol{w}_h^{k+1},\boldsymbol{\varphi}_h) + l_h(\boldsymbol{w}_h^{k+1},\boldsymbol{\varphi}_h) = 0, \quad \forall\ \boldsymbol{\varphi}_h \in \boldsymbol{S}_h, \tag{10}$$

for all $k = 0, 1, \ldots$. Equation (10) is nonlinear with respect to the unknown $\boldsymbol{w}_h^{k+1}$, therefore we linearize the scheme.

In the convective form, we linearized the interior terms using the homogeneity (4) as $\boldsymbol{f}(\boldsymbol{w}_{\mathbf{h}}^{\mathbf{k+1}}) \approx \mathbb{A}(\boldsymbol{w}_{\mathbf{h}}^{\mathbf{k}})\boldsymbol{w}_{\mathbf{h}}^{\mathbf{k+1}}$. In the boundary terms, we use the Vijayasundaram numerical flux (9) and linearize by taking the matrices $\mathbb{P}^{+}$ and $\mathbb{P}^{-}$ at t_k. Thus we get the linearized convective form

$$\begin{aligned}\tilde{b}_h(\boldsymbol{w}_h^k,\boldsymbol{w}_h^{k+1},\boldsymbol{\varphi}_h) = &-\sum_{K\in\mathcal{T}_h}\int_K \mathbb{A}(\boldsymbol{w}_h^k)\boldsymbol{w}_h^{k+1}\cdot\frac{\partial\boldsymbol{\varphi}_h}{\partial x}\,\mathrm{d}x\\ &+\int_{\mathcal{F}_h}\left(\mathbb{P}^{+}(\langle\boldsymbol{w}_h^k\rangle,n)\boldsymbol{w}_h^{k+1,(L)} + \mathbb{P}^{-}(\langle\boldsymbol{w}_h^k\rangle,n)\boldsymbol{w}_h^{k+1,(R)}\right)\cdot[\boldsymbol{\varphi}_h]\,\mathrm{d}S.\end{aligned}$$

As for the source terms, they also need to be linearized to obtain the approximation $l_h(\boldsymbol{w}_h^{k+1}, \boldsymbol{\varphi}_h) \approx \tilde{l}_h(\boldsymbol{w}_h^k, \boldsymbol{w}_h^{k+1}, \boldsymbol{\varphi}_h)$. The specific form of this linearization will be derived in the following. Collecting all these considerations, we obtain the semi-implicit DG scheme:

We seek $\boldsymbol{w}_h^k \in \boldsymbol{S}_h, k = 0, 1, \ldots$, such that for all $\boldsymbol{\varphi}_h \in S_h$

$$\Big(\frac{\boldsymbol{w}_h^{k+1} - \boldsymbol{w}_h^k}{\tau_k}, \boldsymbol{\varphi}_h\Big) + \tilde{b}_h(\boldsymbol{w}_h^k, \boldsymbol{w}_h^{k+1}, \boldsymbol{\varphi}_h) + \tilde{l}_h(\boldsymbol{w}_h^k, \boldsymbol{w}_h^{k+1}, \boldsymbol{\varphi}_h) = 0. \tag{11}$$

Equation (11) represents a linear equation for the unknown $\boldsymbol{w}_h^{k+1}$. If we choose a basis of the space $\boldsymbol{S}_h$ consisting of functions whose support is exactly one element, we can rewrite Eq. (11) as a system of linear algebraic equations for the coefficients of $\boldsymbol{w}_h^{k+1}$ in the chosen basis. If $\tilde{l}_h \equiv 0$ (i.e. we solve the Euler equations), by grouping together basis functions with a common supporting element, the structure of the system matrix is block-tridiagonal with lower-left and upper-right corner blocks corresponding to the periodic boundary conditions. Such systems can be efficiently solved e.g. by a direct solver, in our case UMFPACK, [1].

2.3.1 Linearization of the Source Terms l_h

First, it is necessary to rewrite the right-hand side integrals $\mathcal{A}, \mathcal{B}$ in terms of $\boldsymbol{w}$. For $\mathcal{A}$, we obtain

$$\begin{aligned}\mathcal{A} &= \int_{\mathbb{R}} b(|x-y|)\Big(\underbrace{\rho(x,t)}_{w_1(x,t)}\underbrace{\rho(y,t)u(y,t)}_{w_2(y,t)} - \underbrace{\rho(y,t)}_{w_1(y,t)}\underbrace{\rho(x,t)u(x,t)}_{w_2(x,t)}\Big)\,\mathrm{d}y\\ &= \int_{\mathbb{R}} b(|x-y|)\Big(w_1(x,t)w_2(y,t) - w_1(y,t)w_2(x,t)\Big)\,\mathrm{d}y\\ &= \int_{\mathbb{R}} b(|x-y|)\boldsymbol{w}(x,t)\cdot\big(w_2(y,t), -w_1(y,t), 0\big)\,\mathrm{d}y.\end{aligned}$$

Similarly, we can write $\mathcal{B}$ as

$$\begin{aligned}\mathcal{B} &= \int_{\mathbb{R}} b(|x-y|)\Big(\underbrace{\rho(x,t)u(x,t)}_{w_2(x,t)}\underbrace{\rho(y,t)u(y,t)}_{w_2(y,t)} - 2\underbrace{\rho(x,t)}_{w_1(x,t)}\underbrace{E(y,t)}_{w_3(y,t)}\Big)\,\mathrm{d}y\\ &= \int_{\mathbb{R}} b(|x-y|)\Big(w_2(x,t)w_2(y,t) - 2w_1(x,t)w_3(y,t)\Big)\,\mathrm{d}y\\ &= \int_{\mathbb{R}} b(|x-y|)\boldsymbol{w}(x,t)\cdot\big(-2w_3(y,t), w_2(y,t), 0\big)\,\mathrm{d}y.\end{aligned}$$

Therefore, we can rewrite the vector $\boldsymbol{g}(\boldsymbol{w})$ as

$$\boldsymbol{g}(\boldsymbol{w})(x,t) = \lambda \int_{\mathbb{R}} b(|x-y|)\mathbb{U}_2\big(\boldsymbol{w}(y,t)\big)\boldsymbol{w}(x,t)\,\mathrm{d}y, \tag{12}$$

where $\mathbb{U}_2(\boldsymbol{w}) \in \mathbb{R}^{3\times 3}$ is the matrix

$$\mathbb{U}_2(\boldsymbol{w}) = \begin{pmatrix} 0 & 0 & 0 \\ w_2 & -w_1 & 0 \\ -2w_3 & w_2 & 0 \end{pmatrix}.$$

If we approximate $\boldsymbol{w}(x,t) \approx \boldsymbol{w}_h^{k+1}(x)$ and $\boldsymbol{w}(y,t) \approx \boldsymbol{w}_h^k(y)$, we obtain the linearized form

$$\tilde{l}_h(\boldsymbol{w}_h^k, \boldsymbol{w}_h^{k+1}, \boldsymbol{\varphi}_h) = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} b(|x-y|)\mathbb{U}_2\big(\boldsymbol{w}_h^k(y)\big)\mathrm{d}y \right) \boldsymbol{w}_h^{k+1}(x) \cdot \boldsymbol{\varphi}_h(x)\,\mathrm{d}x. \tag{13}$$

For a basis for $\boldsymbol{S}_h$ formed by functions whose support is only one element, adding (13) does not change the structure of the system matrix, since it contributes only to the block-diagonal. This is important, since other expressions than (12) are possible, however they lead to a full system matrix.

The computation of (13) is very time consuming due to the nonlocal nature. Even if the basis functions of $\boldsymbol{S}_h$ are local, in order to evaluate $\tilde{l}_h$, we must compute the inner integral $\int_{\mathbb{R}} b(|x-y|)\mathbb{U}_2\big(\boldsymbol{w}_h^k(y)\big)\mathrm{d}y$, which is expensive due to the slow decay of the function $b(|x-y|)$. We note that in our implementation, we do not compute this integral over the whole of $\mathbb{R}$, but only over one periodically taken copy of Ω centered at point x.

3 Numerical Experiment

In this numerical experiment, we set the initial density to have a Gaussian distribution $\rho(x) = \exp(-10(x-0.5)^2)$. The temperature is taken constant, $T = 10$, and the velocity is given by $u(x) = -\sin(2\pi x)$. We used 400 piecewise quadratic elements. We observed the formation of a sharp peak in the density, as seen in Fig. 1. Due to the discontinuities in the solution, artificial diffusion was added, as described in [2]. Furthermore, in large regions of Ω, a state close to *vacuum* occurs, i.e. $\rho \approx 0, T \approx 0$. In fact, the minimum density and temperature over Ω seems to decay exponentially, cf. Fig. 2. To avoid this complication, at each time step, the $\boldsymbol{w}_h^k$ was *postprocessed* to avoid the vacuum state. Specifically, if $\rho < \varepsilon$, then set $\rho := \varepsilon$ and recompute the energy, so that $T > \varepsilon$, where $\varepsilon := 10^{-5}$ in our case. A uniform time step $\tau = 10^{-3}$ was chosen as a balance between discretization error in time and computational efficiency.

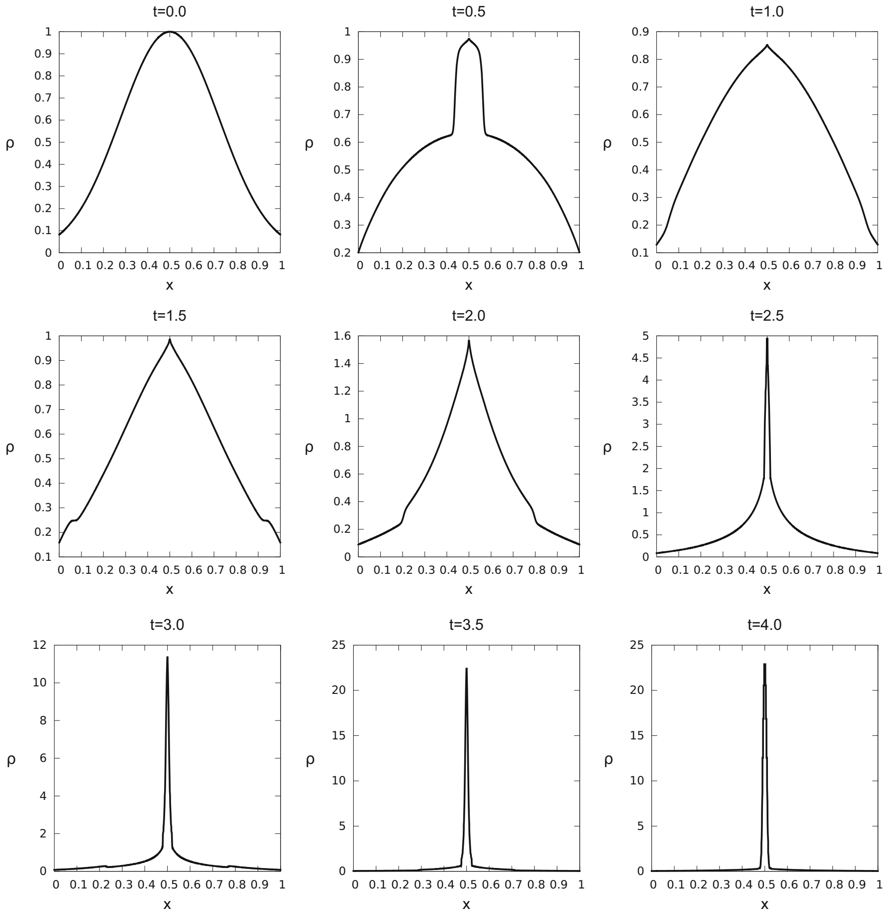


Fig. 1 Time evolution of the density distribution

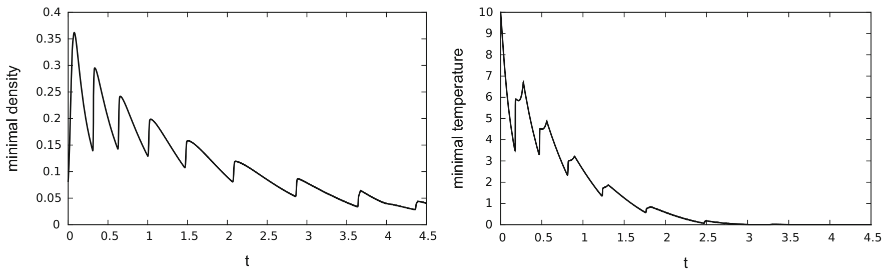


Fig. 2 Time evolution of minimal density and temperature

4 Conclusion

We have presented an efficient numerical method for the solution of a complicated nonlinear and nonlocal version of the compressible Euler equations describing the dynamics of flocks of birds, cf. [4]. To avoid severe time step restrictions and consequently the need to evaluate the expensive nonlocal terms too many times, a semi-implicit discontinuous Galerkin scheme is applied. A suitable treatment of the nonlocal terms is given, which leads to sparse matrices. Shock capturing and postprocessing of vacuum must be added to obtain a stable scheme.

Acknowledgements The research of V. Kučera is supported by the Grant No. P201/13/00522S of the Czech Science Foundation. He is currently a Fulbright visiting scholar at Brown University, Providence, RI, USA, supported by the J. William Fulbright Commission in the Czech Republic. The research of A. Živčáková is supported by the Charles University in Prague, project GA UK No. 758214.

References

1. T.A. Davis, I.S. Duff, A combined unifrontal/multifrontal method for unsymmetric sparse matrices. ACM Trans. Math. Softw. **25**, 1–19 (1999)
2. M. Feistauer, V. Kučera, On a robust discontinuous Galerkin technique for the solution of compressible flow. J. Comput. Phys. **224**, 208–221 (2007)
3. M. Feistauer, J. Felcman, I. Straškraba, *Mathematical and Computational Methods for Compressible Flow* (Clarendon Press, Oxford, 2003)
4. M. Fornasier, J. Haškovec, G. Toscani, Fluid dynamic description of flocking via Povzner-Boltzmann equation. Physica D **240**(1), 21–31 (2011)
5. G. Vijayasundaram, Transonic flow simulation using upstream centered scheme of Godunov type in finite elements. J. Comput. Phys. **63**, 416–433 (1986)

Discontinuous and Enriched Galerkin Methods for Phase-Field Fracture Propagation in Elasticity

Prashant Mital, Thomas Wick, Mary F. Wheeler, and Gergina Pencheva

Abstract In this work, we introduce discontinuous Galerkin and enriched Galerkin formulations for the spatial discretization of phase-field fracture propagation. The nonlinear coupled system is formulated in terms of the Euler-Lagrange equations, which are subject to a crack irreversibility condition. The resulting variational inequality is solved in a quasi-monolithic way in which the irreversibility condition is incorporated with the help of an augmented Lagrangian technique. The relaxed nonlinear system is treated with Newton's method. Numerical results complete the present study.

1 Introduction

Fracture propagation in elasticity, plasticity, and porous media is currently one of the major research topics in mechanical, energy, and environmental engineering. In this paper, we concentrate specifically on fracture propagation in elasticity. We consider a variational approach for brittle fracture introduced in [6], which has been later formulated in terms of a thermodynamically-consistent phase-field technique [8]. In fact, variational and phase-field formulations for fracture are active research areas as attested in recent years, e.g., [1–4, 9, 10]. Our motivations for employing a phase-field model are that fracture nucleation, propagation, kinking, and curvilinear paths are automatically included in the model; post-processing of stress intensity factors and remeshing resolving the crack path are avoided. Furthermore, the underlying equations are based on continuum mechanics principles that can be treated with adaptive Galerkin finite elements.

In this work, we extend existing Galerkin formulations for phase-field fracture with regard to two major aspects:

P. Mital • M.F. Wheeler • Gergina Pencheva
ICES, University of Texas at Austin, Austin, TX 78712, USA
e-mail: mital@ices.utexas.edu; mfw@ices.utexas.edu; gergina@ices.utexas.edu

T. Wick (✉)
RICAM, Austrian Academy of Sciences, Altenberger Str. 69, 4040 Linz, Austria
e-mail: thomas.wick@ricam.oeaw.ac.at

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_20

- Spatial discretization of the displacement field with discontinuous Galerkin (DG) finite elements resulting in NIPG [12] and IIPG methods [5] and an enriched Galerkin (EG) formulation [13];
- Formulation of a quasi-monolithic augmented Lagrangian iteration for the nonlinear coupled displacement-phase-field system.

These frameworks are formulated in Sects. 2, 3 and 4 and are substantiated with numerical tests in Sect. 5.

2 The Phase-Field Fracture Model

We limit our attention to 2-dimensional problems and let $\Omega \in \mathbb{R}^2$, be a smooth, open, connected and bounded set. We denote the L^2 scalar product with $(\cdot,\cdot)$, and assume that the crack $\mathscr{C}$ is a 1-dimensional set, not necessarily connected, contained in Ω. Using the variational/phase-field approach to fracture [3, 6], the crack $\mathscr{C}$ is represented using a continuous phase-field variable $\varphi : \Omega \to [0,1]$. This value of the phase-field variable interpolates between the broken ($\varphi = 0$) and unbroken ($\varphi = 1$) states of the material. The diffusive transition zone between these two states is controlled by a regularization parameter $\varepsilon > 0$. Imposing a crack irreversibility condition $\varphi \leq \varphi^{n-1}$ (where $\varphi^{n-1} := \varphi(t^{n-1})$ denotes the previous time step solution), and further ingredients for a thermodynamically consistent phase-field framework [8] result in the following Euler-Lagrange formulation:

Formulation 1 For the loading steps $n = 1, 2, 3, \ldots$: Find vector-valued displacements and a scalar-valued phase-field variable $\{\mathbf{u}^n, \varphi^n\} := \{\mathbf{u}, \varphi\} \in \{\bar{\mathbf{u}} + V\} \times W$ such that

$$\Big(\big((1-\kappa)\varphi^2 + \kappa\big)\sigma(\mathbf{u}), e(\mathbf{w})\Big) = 0 \quad \forall \mathbf{w} \in V, \tag{1}$$

as well as,

$$\begin{aligned}&(1-\kappa)(\varphi\sigma(\mathbf{u}) : e(\mathbf{u}), \psi - \varphi)\\ &\quad + G_c\Big(-\frac{1}{\varepsilon}(1-\varphi, \psi-\varphi) + \varepsilon(\nabla\varphi, \nabla\psi - \varphi)\Big) \geq 0 \quad \forall \psi \in W_{in} \cap L^{\infty}(\Omega),\end{aligned} \tag{2}$$

where $V := H_0^1(\Omega)$, $W_{in} := \{w \in H^1(\Omega) | w \leq \varphi^{n-1} \leq 1 \text{ a.e. on } \Omega\}$ and $W := H^1(\Omega)$. Furthermore, $\sigma = \sigma(\mathbf{u}) = 2\mu_s e + \lambda_s tr(e) I$ is the stress tensor with $\mu_s, \lambda_s > 0$, and $e(\mathbf{u}) = 0.5(\nabla\mathbf{u} + \nabla\mathbf{u}^T)$ is the linearized strain tensor. The critical energy release rate is $G_c > 0$. The domain is subject to boundary conditions, and we assume $\Gamma_D \neq \emptyset$, with the possibly non-homogeneous and time-dependent Dirichlet boundary conditions $\bar{\mathbf{u}}$. Moreover, κ is a regularization parameter for the elastic energy bounded below by 0, such that $\kappa \ll \varepsilon$, see e.g., [3].

Algorithm 2 Solution algorithm

```
For each time $t^n$:
Let $m = 0$; choose initial $\lambda_m \in L^2(\Omega)$, $\gamma > 0$.
repeat
    Let $k = 0$; choose initial $\tilde{U}_k \in V_h \times W_h$.
    repeat
        Find $\delta U_k$ solving $A'(U_k)(\delta U_k, \Psi) = -A(U_k)(\Psi)$
        Update $\tilde{U}_{k+1} \leftarrow \tilde{U}_k + \delta U_k$
        Update $k \leftarrow k + 1$
    until Stopping criterion $\|U_k - U_{k-1}\| \leq \mathrm{TOL}_2$ is satisfied.
    Set $U_{m+1} = (\mathbf{u}_{m+1}, \varphi_{m+1}) = \tilde{U}_k$
    Update $\lambda_{m+1} = \min(0, \lambda_m + \gamma\varphi_{m+1}) + (\lambda_m + \gamma(\varphi_{m+1} - \varphi^{n-1}))^+$
    Update $m \leftarrow m + 1$
until Stopping criterion $\max(\|\mathbf{u}_{m+1} - \mathbf{u}_m\|, \|\lambda_{m+1} - \lambda_m\|) \leq \mathrm{TOL}_i$ is satisfied.
Set $U^n := \{\mathbf{u}^n, \varphi^n\} = \{\mathbf{u}_{m+1}, \varphi_{m+1}\}$
```

To treat crack irreversibility, we use the augmented-Lagrangian formulation described in [14]. To apply this method, we begin by approximating the time derivative $\partial_t\varphi$ using the backward difference

$$\partial_t\varphi \approx \partial_{\Delta t}\varphi = \frac{\varphi - \varphi^{n-1}}{\Delta t} \quad \Rightarrow \quad \frac{1}{\Delta t}\Big((\lambda + \gamma(\varphi - \varphi^{n-1}))^+\Big), \quad \Delta t = t^n - t^{n-1}.$$

Here, λ and γ are a penalization function and parameter, respectively, and φ^{n-1} is the phase-field solution at the previous time step. Moreover, $(x)^+ := \max\{0, x\}$.

3 A Quasi-monolithic Incremental Formulation

We choose a quasi-monolithic approach [7] as this reduces algorithmic complexity and has been demonstrated to be numerically robust and efficient when φ is replaced by the extrapolation $\tilde{\varphi}$ in the first term of Formulation 2. The reason for choosing $\tilde{\varphi}$ is the need to circumvent the non-convexity of the underlying energy functional.

Formulation 2 For $n = 1, 2, 3, \ldots$: Find $U^n := U := \{\mathbf{u}, \varphi\} \in \{\bar{\mathbf{u}} + V\} \times W$, where $V := H_0^1(\Omega)$ and $W := H^1(\Omega)$, such that

$$A(U)(\Psi) = 0 \quad \forall \Psi := \{\mathbf{w}, \psi\} \in V \times W, \tag{3}$$

where $A(\cdot)(\cdot)$ is the following semi-linear form

$$\begin{aligned} A(U)(\Psi) &= \Big(\big((1-\kappa)\tilde{\varphi}^2 + \kappa\big)\sigma(\mathbf{u}), e(\mathbf{w})\Big) + (1-\kappa)(\varphi\sigma(\mathbf{u}) : e(\mathbf{u}), \psi) \\ &+ G_c\Big(-\frac{1}{\varepsilon}(1-\varphi, \psi) + \varepsilon(\nabla\varphi, \nabla\psi)\Big) + \frac{1}{\Delta t}\Big((\lambda + \gamma(\varphi - \varphi^{n-1}))^+, \psi\Big). \end{aligned} \tag{4}$$

Here $\tilde{\varphi}$ is a linear extrapolation of time-lagged φ, i.e. $\varphi \approx \tilde{\varphi} := \tilde{\varphi}(\varphi^{n-1}, \varphi^{n-2})$, with $\varphi^{n-1}, \varphi^{n-2}$ denoting solutions to previous time steps. Solving the nonlinear variational problem (4) is performed with Newton's method and line search backtracking. The resulting solution algorithm is outlined in Algorithm 2.

4 Spatial Discretization with DG and EG

In this section we establish key notations for DG and EG followed by the mathematical statement of the discrete variational forms. On a conforming subdivision $\mathscr{E}_h$ of a polygonal domain Ω subdivided into elements E we define the discontinuous finite element subspace to be

$$\mathscr{D}_k(\mathscr{E}_h) = \{v \in L^2(\Omega) : \forall E \in \mathscr{E}_h, v|_E \in \mathbb{P}_k(E)\}, \tag{5}$$

where $\mathbb{P}_k(E)$ denotes the space of piecewise polynomials of total degree less than or equal to k on E. We also define the space of CG approximating polynomials enriched with discontinuous piecewise constants

$$\mathscr{D}_k^{C0}(\mathscr{E}_h) := \mathscr{D}_k^C(\mathscr{E}_h) \cup \mathscr{D}_0(\mathscr{E}_h). \tag{6}$$

Here $\mathscr{D}_k^C(\mathscr{E}_h)$ is the CG approximating space defined as

$$\mathscr{D}_k^C(\mathscr{E}_h) = \{v \in C(\Omega) : \forall E \in \mathscr{E}_h, v|_E \in \mathbb{P}_k^C(E), v|_{\Gamma_D} = 0\}, \tag{7}$$

where $\mathbb{P}_k^C(E)$ denotes the space of continuous piecewise polynomials of total degree less than or equal to k on E.

In order to describe the vector-valued displacements, we consider the spaces of vector functions that generalize (5) and (6): $\mathscr{D}_k(\mathscr{E}_h) = (\mathscr{D}_k(\mathscr{E}_h))^d$, $\mathscr{D}_k^{C0}(\mathscr{E}_h) = (\mathscr{D}_k^{C0}(\mathscr{E}_h))^d$, where d is the number of spatial dimensions. We note that the functions in $\mathscr{D}_k(\mathscr{E}_h)$ and $\mathscr{D}_k^{C0}(\mathscr{E}_h)$ are discontinuous along the edges (or faces) of the mesh.

Now, consider two neighboring elements E_1^e and E_2^e that share a common side e. Naturally then, there are two traces of $w \in \mathscr{D}_k(\mathscr{E}_h)$ along e (see Fig. 1). We consider $\mathbf{n}_e$ to be the normal vector associated with e to be oriented from E_1^e to E_2^e and define: $\{\mathbf{w}\} = \frac{1}{2}(\mathbf{w}|_{E_1^e}) + \frac{1}{2}(\mathbf{w}|_{E_2^e})$, $\quad [\mathbf{w}] = (\mathbf{w}|_{E_1^e}) - (\mathbf{w}|_{E_2^e}) \quad \forall e = \partial E_1^e \cap \partial E_2^e$. We extend this definition to elements on the boundary $\partial\Omega$ as: $\{\mathbf{w}\} = [\mathbf{w}] = (\mathbf{w}|_{E_1^e}) \quad \forall e = \partial E_1^e \cap \partial\Omega$. Further, we denote by $|e|$ the length of an edge e in $d = 2$. We now state the equations corresponding to a discontinuous spatial discretization directly from inspection of the monolithic formulation (4) and the DG-scheme for pure linear elasticity, e.g., [11]. We pursue a discontinuous representation of the displacement variable $\mathbf{u}$ only, recognizing that the regularization in the case of the phase-field variable φ enforces its continuity across the crack.

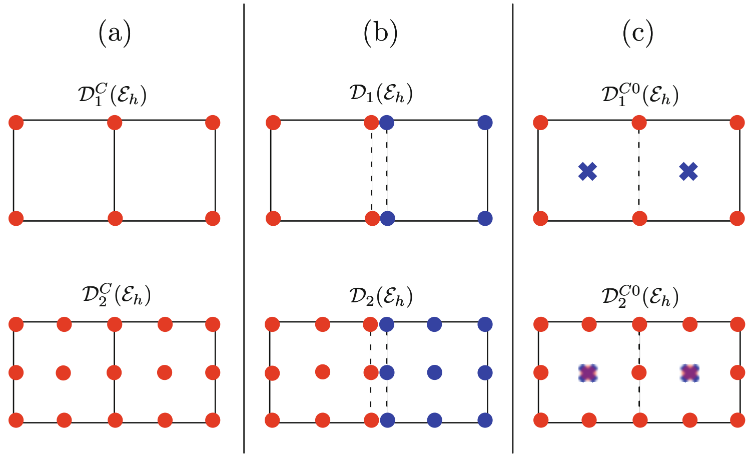

(a)
$\mathcal{D}_1^C(\mathcal{E}_h)$
$\mathcal{D}_2^C(\mathcal{E}_h)$
(b)
$\mathcal{D}_1(\mathcal{E}_h)$
$\mathcal{D}_2(\mathcal{E}_h)$
(c)
$\mathcal{D}_1^{C0}(\mathcal{E}_h)$
$\mathcal{D}_2^{C0}(\mathcal{E}_h)$

Fig. 1 Support points for bilinear and biquadratic basis functions. (**a**) CG: all support points in *red*. (**b**) DG: support points for *left* element in *red*, for *right* element in *blue*. The common edge has two sets of support points – one from each element. (**c**) EG: support points from CG approximating space in *red*, piecewise constants in *blue*. Only the piecewise constant degree of freedom is discontinuous across the common edge

We augment (4) with the jump and penalization terms to define the discrete incremental semi-linear form

$$
\begin{aligned}
A(U_h)(\Psi_h) &= \sum_{E\in\mathscr{E}_h}\int_E\Big(((1-\kappa)\tilde{\varphi}^2+\kappa)\sigma(\mathbf{u}), e(\mathbf{w})\Big) \\
&- \sum_{e\in\Gamma_h\cup\Gamma_D}\int_e \{((1-\kappa)\tilde{\varphi}^2+\kappa)\sigma(\mathbf{u})\cdot\mathbf{n}_e\}[\mathbf{w}] + \eta\sum_{e\in\Gamma_h}\int_e\{((1-\kappa)\tilde{\varphi}^2+\kappa)\sigma(\mathbf{w})\cdot\mathbf{n}_e\}[\mathbf{u}] \\
&+ \eta\sum_{e\in\Gamma_D}\int_e\Big(((1-\kappa)\tilde{\varphi}^2+\kappa)\sigma(\mathbf{w})\cdot\mathbf{n}_e\Big)(\mathbf{u}-\mathbf{g}_D) + \sum_{e\in\Gamma_h}\frac{\delta_e}{|e|^\beta}\int_e[\mathbf{u}][\mathbf{w}] + \sum_{e\in\Gamma_D}\frac{\delta_e}{|e|^\beta}\int_e(\mathbf{u}-\mathbf{g}_D)\mathbf{w} \\
&+ (1-\kappa)(\varphi\sigma(\mathbf{u}):e(\mathbf{u}),\psi) + G_c\Big(-\frac{1}{\varepsilon}(1-\varphi,\psi)+\varepsilon(\nabla\varphi,\nabla\psi)\Big) + \frac{1}{\Delta t}\Big((\lambda+\gamma(\varphi-\varphi^{n-1}))^+,\psi\Big),
\end{aligned}
\tag{8}
$$

where $\eta = 1$ (NIPG) or $\eta = 0$ (IIPG); and $\delta_e > 0$ and $\beta > 0$ (here $\beta = 1$) are the DG-penalization and superpenalization parameters, respectively. The DG-CG variational problem reads: Find $U_h := \{\mathbf{u}, \varphi\} \in \{\bar{\mathbf{u}}_h + V_h^{DG}\} \times W_h^{CG}$ such that

$$
A(U_h)(\Psi_h) = 0 \quad \forall \Psi_h := \{\mathbf{w}, \psi\} \in V_h^{DG} \times W_h^{CG}. \tag{9}
$$

The EG-CG variational problem reads: Find $U_h := \{\mathbf{u}, \varphi\} \in \{\bar{\mathbf{u}}_h + V_h^{EG}\} \times W_h^{CG}$ such that

$$A(U_h)(\Psi_h) = 0 \quad \forall \Psi_h := \{\mathbf{w}, \psi\} \in V_h^{EG} \times W_h^{CG}. \tag{10}$$

The test and trial spaces are $V_h^{DG} := [\mathscr{D}_k(\mathscr{E}_h)]^2, V_h^{EG} := [\mathscr{D}_k^{C0}(\mathscr{E}_h)]^2, W_h^{CG} := \mathscr{D}_k^{C}(\mathscr{E}_h)$. Our formulation enforces both Dirichlet and Neumann boundary conditions weakly. The use of homogeneous Neumann boundary conditions for $\mathbf{u}$ and φ results in a formulation exclusively dependent on Γ_D. The directional derivative of (8) needed for the Newton iterations is computed analytically.

5 A Numerical Test: Single Edge Notched Tension

The single edge notched tension test is a widely used experimental methodology used to characterize the fracture toughness of various materials in plane-strain.

We consider a square plate with a horizontal notch placed at half-height, running from the right outer surface to the center of the specimen. The plate is subject to zero displacement boundary conditions on the bottom surface, and time-dependent displacement on the top surface. The left and right surfaces are considered to be traction-free. The problem setup is shown in Fig. 2. The material parameters are chosen as $\lambda = 121.1538\,\text{kN/mm}^2$, $\mu = 80.7692\,\text{kN/mm}^2$ and $G_c = 2.7 \times 10^{-3}\,\text{kN/mm}$. The displacement boundary condition on the top surface is taken to be $\bar{u}_y(t) = t\bar{\alpha}$ with $\bar{\alpha} = 1\,\text{mm/s}$. The expected response of this test is the build-up of the stress concentration in the vicinity of the crack-tip, followed by unstable, catastrophic crack growth.

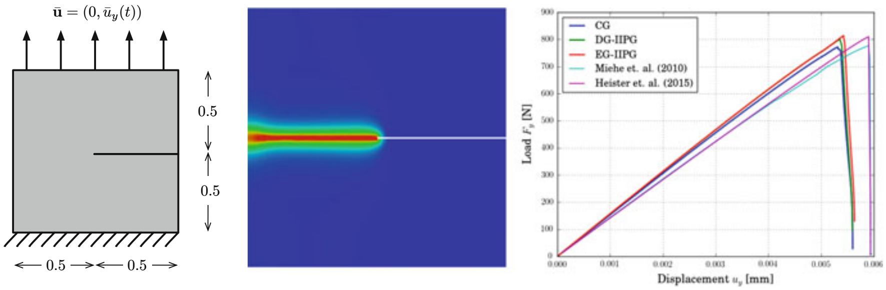


Fig. 2 Schematic of the single-edge-notched tension test (*left*), the final phase-field crack pattern at $T = 6.6 \times 10^{-3}$ s (*middle*), and comparison of load-displacement curves from our monolithic scheme with CG, DG-IIPG and EG-IIPG against results reported by Miehe et al. [9] and Heister et al. [7]

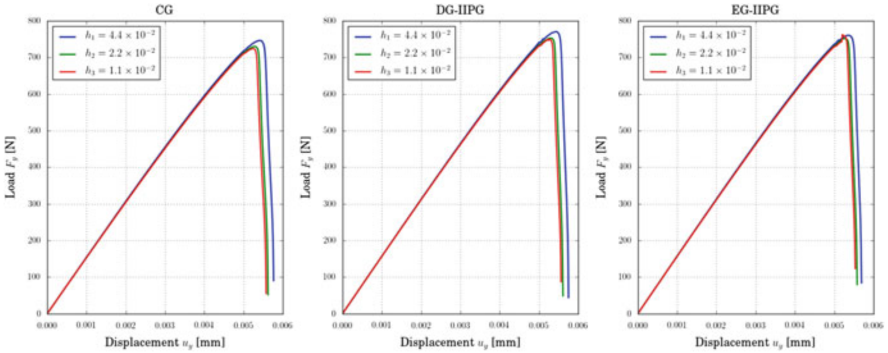


Fig. 3 We fix κ and ε on the coarsest mesh level and vary h. *Left*: CG. *Middle*: DG-IIPG. *Right*: EG-IIPG. Spatial convergence is observed for all schemes

Our first objective is to study h-convergence for fixed ε. We choose $\Delta t = 1.0 \times 10^{-4}$ s for the first 50 loading steps, after which $\Delta t = 1.0 \times 10^{-5}$ s. This adaptivity in the time step is necessary to capture the rapid movement of the crack tip. We choose $\varepsilon = 4.4 \times 10^{-2}$ mm, $\kappa = 1.0 \times 10^{-12}$, and run our code for $h_1 = 4.4 \times 10^{-2}$ mm, $h_2 = 2.2 \times 10^{-2}$ mm, and $h_3 = 1.1 \times 10^{-2}$ mm. We evaluate the surface load vector on the top surface of Ω as $\tau = (F_x, F_y) = \int_{\partial\Omega_{\text{top}}} \sigma(\mathbf{u})\mathbf{n}ds$. In this example, we are particularly interested in F_y. Our findings for the surface load evolution with varying h are shown in Fig. 3 for the IIPG flavors of EG and DG. It is observed that our approach is stable with spatial mesh refinement, and that our solution converges as we use finer meshes. Comparison to literature values are displayed in Fig. 2 at right.

Results obtained from DG-NIPG and EG-NIPG are very similar and are therefore not presented here. The SIPG method ($\eta = -1$) yields unsatisfactory findings, which are not shown in this work.

With the results of our scheme duly validated, we proceed to study the relative efficiency of the schemes by comparing the number of Newton iterations taken by each of them to converge. We first investigate the variation in the number of Newton steps taken with the penalization parameter δ_e. Note that when we multiply Equations (8) throughout by Δt, our effective penalization of the jump becomes $\delta_e \Delta t$. This is an important detail that cannot be overlooked while using DG/EG for the phase-field equations because for instance, using $\delta_e = 10^5$ with $\Delta t = 1 \times 10^{-5}$ s gives an effective penalization of $\delta_e \Delta t = 1$ which is not sufficiently large and produces spurious results. In the case of an adaptive time step size (we usually take $\Delta t = 1 \times 10^{-4}$ s for the first 50 steps, and a smaller time step thereafter), the product $\delta_e \Delta t$ is reported for the smaller time step. We vary the values of the effective penalization and plot the cumulative number of Newton steps as a function of time for $h = 1.1 \times 10^{-2}$ mm, $\varepsilon = 2\,h$[mm], and $\kappa = 1.0 \times 10^{-12}$. The results of this study with IIPG and NIPG are shown in Fig. 4. In Fig. 5, we observe that DG and EG

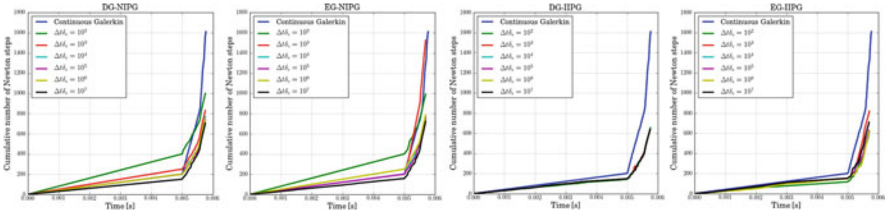


Fig. 4 Newton convergence performance with $h = 1.1 \times 10^{-2}$ mm, $\Delta t = \{10^{-4}, t < 0.005; 10^{-5}, t \geq 0.005\}$, $\varepsilon = 2h$[mm], and $\kappa = 10^{-12}$ and varying penalty δ_e. *Left*: DG-NIPG. *Right*: EG-NIPG. *Left*: DG-IIPG. *Right*: EG-IIPG. Convergence is faster for higher values of penalization

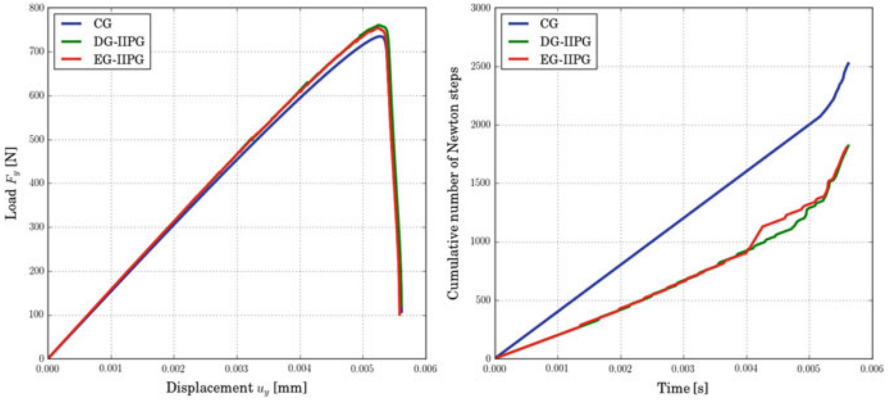


Fig. 5 Single edge notched tension test results using CG, DG-IIPG and EG-IIPG. *Left*: load vs. displacement curve. *Right*: Newton convergence performance for constant $\Delta t = 10^{-5}$ s

schemes take much fewer Newton iterations to converge than CG especially after the onset of crack growth (approximately $t = 0.0055$ s).

For a better comparison of the efficiency, we run a test with the same physical parameters as above, but with a uniform time step of $\Delta t = 10^{-5}$ s throughout. The motivation is to suppress the effect of adaptive time stepping on the Newton performance and to give an unbiased comparison. Since the computational burden with such a simulation is significant, we only consider the IIPG case with $\delta_e \Delta t = 10^2$. These results are shown in Fig. 5. As we can see, DG and EG take roughly the same number of Newton iterations (1800) while CG takes significantly more (2520). We also observe that the load-displacement curves for all three methods are in reasonable agreement. Hence, we can conclusively state that the Newton method converges in fewer iterations for the DG and EG schemes than for the CG scheme. Furthermore by inspecting Fig. 1, we see that EG has significantly fewer degrees of freedom than DG.

References

1. M. Ambati, T. Gerasimov, L. De Lorenzis, A review on phase-field models of brittle fracture and a new fast hybrid formulation. Comput. Mech. **55**(2), 383–405 (2015)
2. M.J. Borden, C.V. Verhoosel, M.A. Scott, T.J. Hughes, C.M. Landis, A phase-field description of dynamic brittle fracture. Comput. Methods Appl. Mech. Eng. **217**, 77–95 (2012)
3. B. Bourdin, G.A. Francfort, J.J. Marigo, The variational approach to fracture. J. Elast. **91**(1–3), 5–148 (2008)
4. S. Burke, C. Ortner, E. Süli, An adaptive finite element approximation of a variational model of brittle fracture. SIAM J. Numer. Anal. **48**(3), 980–1012 (2010)
5. C. Dawson, S. Sun, M.F. Wheeler, Compatible algorithms for coupled flow and transport. Comput. Methods Appl. Mech. Eng. **193**(23), 2565–2580 (2004)
6. G.A. Francfort, J.-J. Marigo, Revisiting brittle fracture as an energy minimization problem. J. Mech. Phys. Solids **46**(8), 1319–1342 (1998)
7. T. Heister, M.F. Wheeler, T. Wick, A primal-dual active set method and predictor-corrector mesh adaptivity for computing fracture propagation using a phase-field approach. Comput. Methods Appl. Mech. Eng. **290**, 466–495 (2015)
8. C. Miehe, F. Welschinger, M. Hofacker, Thermodynamically consistent phase-field models of fracture: variational principles and multi-field fe implementations. Int. J. Numer. Methods in Eng. **83**(10), 1273–1311 (2010)
9. C. Miehe, M. Hofacker, F. Welschinger, A phase-field model for rate-independent crack propagation: robust algorithmic implementation based on operator splits. Comput. Methods Appl. Mech. Eng. **199**(45), 2765–2778 (2010)
10. A. Mikelić, M.F. Wheeler, T. Wick, A quasi-static phase-field approach to pressurized fractures. Nonlinearity **28**(5), 1371 (2015)
11. B. Rivière, *Discontinuous Galerkin Methods for Solving Elliptic and Parabolic Equations: Theory and Implementation* (SIAM, Philadelphia, 2008)
12. B. Rivière, M.F. Wheeler, K. Banaś et al., Discontinuous Galerkin method applied to a single phase flow in porous media. Comput. Geosci. **4**(4), 337–349 (2000)
13. S. Sun, J. Liu, A locally conservative finite element method based on piecewise constant enrichment of the continuous Galerkin method. SIAM J. Sci. Comput. **31**(4), 2528–2548 (2009)
14. M.F. Wheeler, T. Wick, W. Wollner, An augmented-lagrangian method for the phase-field approach for pressurized fractures. Comput. Methods Appl. Mech. Eng. **271**, 69–85 (2014)

Numerical Method Based on DGM for Solving the System of Equations Describing Motion of Viscoelastic Fluid with Memory

Ivan Soukup

Abstract We present a numerical method for the solution of integro-differential equations describing motion of an incompressible viscoelastic fluid with memory. In particular, the system of equations consists of the momentum conservation equation with the Cauchy stress tensor divided in a viscous and an elastic parts which depend non-linearly on the symmetric part of velocity gradient and non-linearly on the past values of the Finger strain tensor, respectively. The momentum conservation equation is completed with system of equations that describes relation between the velocity gradient and the Finger strain tensor. The method is based on a discontinuous Galerkin method in the spatial variables and the BDF methods in the time variables.

1 Introduction

Viscoelastic fluids appear in many aspects of life – from the nature to man-made materials. The need to understand the behaviour of such fluids is evident. One of the way to achieve a better understanding is to apply simulation tools that involves numerical methods for solving PDEs.

Mathematically, the system of equations describing an incompressible and an isothermal viscoelastic fluid motion is expressed by the continuity and momentum equations

$$\nabla \cdot \boldsymbol{v} = 0, \quad \frac{D\boldsymbol{v}}{Dt} = -\nabla\pi + \nabla \cdot \boldsymbol{\tau} + \boldsymbol{f},$$

where $\frac{D}{Dt}$ denotes the material derivative, $\nabla\cdot$ and ∇ denote the divergence and the gradient operators, respectively. The velocity is denoted by $\boldsymbol{v}$, π stands for the pressure, $\boldsymbol{\tau}$ denotes the stress tensor and $\boldsymbol{f}$ represents the external body force. In

I. Soukup (✉)
Charles University in Prague, Sokolovská 83, Prague, Czeh Republic
e-mail: soukup@karlin.mff.cuni.cz

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_21

order to complete the above system of equations, we have to specify the constitutive law.

Basically, there are two classes of constitutive equations for viscoelastic fluids, differential and integral. We focus here on the integral type since it is more general and usually the most physically precise approach. Most of the integral models we are interested in have separated viscous and elastic part of the stress tensor, i.e.

$$\boldsymbol{\tau} = \boldsymbol{\sigma}_n + \boldsymbol{\sigma}_e.$$

We work in the Eulerian framework and assume the following description of the Newtonian component of the stress tensor

$$\boldsymbol{\sigma}_n = \mu \mathbb{D},$$

where $\mathbb{D}$ denotes the symmetric part of the velocity gradient and μ is the viscosity constant. Further, the elastic part is assumed to be in the form

$$\boldsymbol{\sigma}_e = \int_{-\infty}^{t} \mathscr{G}(t-s)\mathscr{H}(\mathscr{B}'(x,t,s))ds,$$

where $\mathscr{G}$ denotes a so-called memory function, $\mathscr{H}$ is in general non-linear tensor operator and $\mathscr{B}'$ stands for the Finger strain tensor. Let us remind the definition of the Finger strain tensor. Consider a moving fluid particle, that has a position vector $\boldsymbol{x}$ at the present time t and had a position vector $\boldsymbol{x}'$ at some past time s. The deformation gradient tensor $\mathbb{E}$ is defined by $E_{ij} = \frac{\partial x_i}{\partial x_j{}'}$, $i,j = 1,\ldots,d$, $(d = 2,3)$ and expresses displacement of the particle moving from the point $\boldsymbol{x}'$ to $\boldsymbol{x}$. Then the Finger strain tensor is defined by $\mathscr{B}'(\cdot,t,s) = \mathbb{E}(\cdot,t,s) \cdot \mathbb{E}^T(\cdot,t,s)$. Let us note that the Finger strain tensor can be understood as a field that measures the deformation of the fluid element currently (at time t) present at the position x, with respect to the reference time s somewhere in the past. The time evolution of the Finger strain tensor is governed by the equation

$$\frac{D\mathscr{B}'}{Dt} = \nabla \boldsymbol{v} \cdot \mathscr{B}' + \mathscr{B}' \cdot \nabla \boldsymbol{v}^T.$$

Let us note that this general framework covers most of the integral models like Oldroyd-B, Doi Edwards, K-BKZ, Rivlin-Sawyers, Wagner, PSM and many others. In the development of the numerical method we follow mainly the work [2] where the concept of deformation fields is presented. Our method is based on a discontinuous Galerkin method in the spatial variables and BDF methods in time variables. That is in contrast with standard FEM methods and TDG/BDF used in [2] and others like [1, 3, 6] or [4]. The implementation is carried out in FEniCS environment.

2 Problem Formulation

Let $\Omega \subset \mathbb{R}^d$, $d = 2, 3$ be a bounded domain and $T > 0$. We set $Q_T = \Omega \times [0, T]$ and by $\partial\Omega$ we denote the boundary of Ω which consists of two disjoint parts, $\partial\Omega = \Gamma_D \cup \Gamma_N$. Following [2] we introduce the age $\tau = t - s$ as a new independent variable instead of s and define the Deformation field tensor

$$\mathscr{B}(x, t, \tau) \equiv \mathscr{B}'(x, t, t - \tau).$$

All together we are interested in finding $(\boldsymbol{v}, \pi, \mathscr{B})$ such that

$$\begin{aligned} \nabla \cdot \boldsymbol{v} = 0 \text{ in } Q_T \\ \frac{\partial \boldsymbol{v}}{\partial t} + \boldsymbol{v} \cdot \nabla \boldsymbol{v} = -\nabla \pi + \boldsymbol{f} + \nabla \cdot \left(\mu \mathbb{D} + \int_0^{+\infty} \mathscr{G}(\tau) \mathscr{H}(\mathscr{B}(x, t, \tau)) d\tau \right) \text{ in } Q_T \\ \frac{\partial \mathscr{B}}{\partial t} + \frac{\partial \mathscr{B}}{\partial \tau} + \boldsymbol{v} \cdot \nabla \mathscr{B} = \nabla \boldsymbol{v} \cdot \mathscr{B} + \mathscr{B} \cdot \nabla \boldsymbol{v}^T, \quad x \in \Omega, t \in [0, T], \tau \in [0, +\infty). \end{aligned}$$

We complete this system with the initial conditions $\boldsymbol{v}(x, 0) = \boldsymbol{v}_0(x)$ in Ω, $\mathscr{B}(x, 0, \tau) = \mathscr{B}_{old}(x, \tau)$ in $\Omega \times [0, +\infty)$ and boundary conditions $\boldsymbol{v} = \boldsymbol{v}_D$ on $\Gamma_D \times (0, T)$, $\mu \mathbb{D} \boldsymbol{n} - \pi \mathbb{I} = \boldsymbol{0}$ on $\Gamma_N \times (0, T)$ and $\mathscr{B}(x, t, 0) = \mathbb{I}$ in Q_T, where $\mathscr{B}_{old}$ is a given function representing all the past deformations of the flow.

3 Discretization

The numerical scheme relies heavily on the use of the backward Euler time discretization of Navier-Stokes equations and the forward Euler time discretization of the equations describing the time evolution of the Deformation field tensor. This semi-implicit approach allows us at first to evaluate the Deformation field tensor and afterwards to compute separately a new velocity field.

Moreover, the spatial discretization of the Navier-Stokes equations is based on the IIPG discretization with the use of upwinding scheme (see [5]). The spatial discretization of the evolution equation of the Deformation field tensor is based on a DG approach with the upwinding as well and the discretization with respect to the age variable is carried through by BDF-2 method.

3.1 Finite-Dimensional Spaces and Subdivision of Time Intervals

Let us start with spatial discretization and let $\mathscr{T}_h$ be a regular subdivision of Ω. We seek an approximation of the velocity, the pressure and the Deformation field tensor in the finite-dimensional spaces

$$\begin{aligned}
\mathscr{V}_{hp} &= \{\boldsymbol{v} \in L^2(\Omega)^d : \forall K \in \mathscr{T}_h, \boldsymbol{v} \in (\mathscr{P}_p(K))^d\},\\
\mathscr{R}_{hq} &= \{r \in L^2(\Omega) : \forall K \in \mathscr{T}_h, r \in \mathscr{P}_q(K)\},\\
\mathscr{Q}_{hp} &= \{\mathscr{B} \in L^2(\Omega)^{d\times d} : \forall K \in \mathscr{T}_h, \mathscr{B} \in (\mathscr{P}_p(K))^{d\times d}\},
\end{aligned}$$

where $\mathscr{P}_p(K)$ and $\mathscr{P}_q(K)$ denotes the space of all polynomials on K of degree $\leq p$ and $\leq q$, respectively.

Let $0 = t_0 < t_1 < \ldots < t_N = T$ be a partition of the interval $(0, T)$ and $\Delta t_k = t_k - t_{k-1}$, $k = 1, \ldots, N$, and let $0 = \tau_1 < \tau_2 < \ldots < \tau_{N_\tau}$ be a partition of $(0, \tau_{N_\tau})$, where $\tau_{N_\tau} \ll \infty$.

3.2 Discretization of the Navier-Stokes Equations

Let us approximate $\boldsymbol{\sigma}_e$ in the following way

$$\begin{aligned}
\int_0^{+\infty} \mathscr{G}(\tau)\mathscr{H}(\mathscr{B}(x,t,\tau))d\tau &\approx \sum_{k=1}^{N_\tau} \omega_k \mathscr{H}(\mathscr{B}(x,t,\tau_k)),\\
\omega_k &= \int_{\tau_k}^{\tau_{k+1}} \mathscr{G}(\tau)d\tau, \qquad k = 1,\ldots,N_\tau - 1,\\
\omega_{N_\tau} &= \int_{\tau_{N_\tau}}^{\infty} \mathscr{G}(\tau)d\tau.
\end{aligned}$$

We denote $\boldsymbol{n}$ the unit normal vector for $\Gamma \subset \partial K$, $K \in \mathscr{T}_h$, and we recall the usual notation for jumps and averages

$$\text{for}\, x \in \Gamma : \boldsymbol{v}_\Gamma^L(x) = \lim_{\varepsilon\to 0,\varepsilon>0} \boldsymbol{v}(x + \varepsilon\boldsymbol{n}), \boldsymbol{v}_\Gamma^R(x) = \lim_{\varepsilon\to 0,\varepsilon>0} \boldsymbol{v}(x - \varepsilon\boldsymbol{n}),$$
$$[\boldsymbol{v}] = \boldsymbol{v}^L - \boldsymbol{v}^R, \quad \langle\boldsymbol{v}\rangle = \tfrac{1}{2}(\boldsymbol{v}^L + \boldsymbol{v}^R).$$

Now, we define for all $\boldsymbol{v}, \boldsymbol{w}, \boldsymbol{\varphi} \in \mathscr{V}_{hp}$, $\pi, r \in \mathscr{R}_{hq}$ and $\sigma > 0$ the following forms

$$\begin{aligned} A_h(\boldsymbol{v}, \boldsymbol{\varphi}) =&\ \mu a_h(\boldsymbol{v}, \boldsymbol{\varphi}) + J_h^\sigma(\boldsymbol{v}, \boldsymbol{\varphi}) \\ a_h(\boldsymbol{v}, \boldsymbol{\varphi}) =&\ \sum_{K \in \mathscr{T}_h} \int_K \mathbb{D}(\boldsymbol{v})\mathbb{D}(\boldsymbol{\varphi}) dx - \sum_{\Gamma \in \Gamma^{ID}} \int_\Gamma \langle \mathbb{D}(\boldsymbol{v}) \rangle \boldsymbol{n} \, [\boldsymbol{\varphi}] \, dS \\ J_h^\sigma(\boldsymbol{v}, \boldsymbol{\varphi}) =&\ \sum_{\Gamma \in \Gamma^{ID}} \int_\Gamma \sigma \, [\boldsymbol{v}] \, [\boldsymbol{\varphi}] \, dS \\ b_h(\pi, \boldsymbol{\varphi}) =&\ -\sum_{K \in \mathscr{T}_h} \int_K \pi \operatorname{div}(\boldsymbol{\varphi}) dx + \sum_{\Gamma \in \Gamma^{ID}} \int_\Gamma \langle \pi \boldsymbol{n} \rangle \, [\boldsymbol{\varphi}] \, dS \\ c_h(\boldsymbol{v}, \boldsymbol{w}, \boldsymbol{\varphi}) =&\ -\frac{1}{2} \sum_{K \in \mathscr{T}_h} \int_K \boldsymbol{v}\boldsymbol{w} \nabla \boldsymbol{\varphi} dx + \sum_{\Gamma \in \Gamma^{IDN}} \int_\Gamma H(\boldsymbol{v}, \boldsymbol{w}, \boldsymbol{n}) \, [\boldsymbol{\varphi}] \, dS, \end{aligned}$$

where

$$\begin{aligned} H(\boldsymbol{v}, \boldsymbol{w}, \boldsymbol{n}) &= \frac{1}{2} \boldsymbol{v}_\Gamma^L \boldsymbol{w}_\Gamma^L \boldsymbol{n}, \quad \text{if } \frac{1}{2}(\boldsymbol{w}_\Gamma^L + \boldsymbol{w}_\Gamma^R)\boldsymbol{n} > 0 \\ &= \frac{1}{2} \boldsymbol{v}_\Gamma^R \boldsymbol{w}_\Gamma^R \boldsymbol{n}, \quad \text{else.} \end{aligned}$$

Moreover, for all $n = 1, \dots, N$

$$\begin{aligned} L^n(\boldsymbol{\varphi}, \mathscr{B}) &= (\boldsymbol{f}(t_n), \boldsymbol{\varphi}) + \sum_{\Gamma \in \Gamma^D} \int_\Gamma \sigma \boldsymbol{\varphi} \boldsymbol{v}_D(t_n) dS \\ &- \sum_{k=1}^{N_\tau} \omega_k \left[\int_\Omega \mathscr{H}(\mathscr{B}(x, t_n, \tau_k)) \mathbb{D}(\boldsymbol{\varphi}) dx - \sum_{\Gamma \in \Gamma^{IDN}} \int_\Gamma \langle \mathscr{H}(\mathscr{B}(x, t_n, \tau_k)) \rangle \boldsymbol{n} [\boldsymbol{\varphi}] dS \right] \\ R(q) &= \sum_{\Gamma \in \Gamma^D} \int_\Gamma q \boldsymbol{v}_D \boldsymbol{n} dS. \end{aligned}$$

Now, by $\boldsymbol{v}_h^0$ we denote the projection of $\boldsymbol{v}_0$ on $\mathscr{V}_{hp}$ and for given $\sigma > 0$, $\mathscr{B}_h \in \mathscr{Q}_{hp}$ and $n \in \{1, \dots, N\}$ we consider $(\boldsymbol{v}_h^n, \pi_h^n) \in \mathscr{V}_{hp} \times \mathscr{R}_{hq}$ the approximate solution of the Navier-Stokes equations at the time t_n, if for all $\boldsymbol{\varphi}_h \in \mathscr{V}_{hp}$ and for all $r_h \in \mathscr{R}_{hq}$, respectively, it holds

$$\begin{aligned} \frac{(\boldsymbol{v}_h^n - \boldsymbol{v}_h^{n-1}, \boldsymbol{\varphi}_h)}{\Delta t_n} + A_h(\boldsymbol{v}_h^n, \boldsymbol{\varphi}_h) + b_h(\pi_h^n, \boldsymbol{\varphi}_h) &+ c_h(\boldsymbol{v}_h^n, \boldsymbol{v}_h^{n-1}, \boldsymbol{\varphi}_h) \\ &= L^n(\boldsymbol{\varphi}_h, \mathscr{B}_h), \\ b_h(r_h, \boldsymbol{v}_h^n) &= R(r_h). \end{aligned}$$

3.3 Discretization of the Evolution Equation for the Finger Strain Tensor

Let us denote $\partial K_n^- \equiv \{x \in \partial K \setminus \partial \Omega : \langle \boldsymbol{v}_h(x, t_n) \rangle \boldsymbol{n} < 0\}$ as the inlet part of the boundary of $K \in \mathscr{T}_h$, $A * B \equiv \{a_{ij} b_{ij}\}_{i,j=1}^d \in \mathbb{R}^{d \times d}$ for all $A, B \in \mathbb{R}^{d \times d}$ and $\left(\underline{\underline{\boldsymbol{u}}}, \underline{\underline{\boldsymbol{v}}}\right)_K = \int_K \underline{\underline{\boldsymbol{u}}} * \underline{\underline{\boldsymbol{v}}} dx$, for all $\underline{\underline{\boldsymbol{u}}}, \underline{\underline{\boldsymbol{v}}} \in \mathscr{Q}_{hp}$.

Further, let α_i, $i = 1, 2, 3$ be the BDF-2 coefficients depending on $\Delta \tau_k = \tau_k - \tau_{k-1}$ and $\Delta \tau_{k-1}$.

Let us also denote $\mathscr{B}_h^{nk}(\cdot) = \mathscr{B}_h(\cdot, t_n, \tau_k) \in \mathscr{Q}_{hp}$ an approximate solution to the evolution equation for the Deformation field tensor at the time t_n and the age τ_k.

Now, using standard DG spatial discretization together with the use of the upwinding scheme, the BDF-2 method for the age discretization and the forward Euler method for the discretization in the time t variable we get the numerical scheme: For given $n = 0, \ldots, N-1$ and given $\boldsymbol{v}_h^n \in \mathscr{V}_{hp}$ find $B_h^{nk} \in \mathscr{Q}_{hp}$ such that for all $\underline{\underline{\boldsymbol{v}}}_h \in \mathscr{Q}_{hp}$ and for all $k = 1, \ldots, N_\tau$ holds

$$
\begin{aligned}
\Delta \tau_k \sum_{K \in \mathscr{T}_h} \left(\mathscr{B}_h^{(n+1)k}, \underline{\underline{\boldsymbol{v}}}_h\right)_K = {} & \Delta \tau_k \sum_{K \in \mathscr{T}_h} \left(\mathscr{B}_h^{nk}, \underline{\underline{\boldsymbol{v}}}_h\right)_K \\
& - \Delta t_{n+1} \sum_{K \in \mathscr{T}_h} \left(\alpha_0 \mathscr{B}_h^{nk} + \alpha_1 \mathscr{B}_h^{n(k-1)} + \alpha_2 \mathscr{B}_h^{n(k-2)}, \underline{\underline{\boldsymbol{v}}}_h\right)_K \\
& - \Delta t_{n+1} \Delta \tau_k \sum_{K \in \mathscr{T}_h} \left(\boldsymbol{v}_h^n \cdot \nabla \mathscr{B}_h^{nk}, \underline{\underline{\boldsymbol{v}}}_h\right)_K \\
& + \Delta t_{n+1} \Delta \tau_k \sum_{K \in \mathscr{T}_h} \left(\nabla \boldsymbol{v}_h^n \cdot \mathscr{B}_h^{nk} - \mathscr{B}_h^{nk} \cdot \nabla (\boldsymbol{v}_h^n)^T, \underline{\underline{\boldsymbol{v}}}_h\right)_K \\
& + \Delta t_{n+1} \Delta \tau_k \sum_{K \in \mathscr{T}_h} \int_{\partial K_n^- \setminus \partial \Omega} \left(\boldsymbol{n} (\boldsymbol{v}_h^n)^T \left[\mathscr{B}_h^{nk}\right]_x\right) * \underline{\underline{\boldsymbol{v}}}_h \, dS,
\end{aligned}
$$

where $\boldsymbol{n}$ denotes the unit normal vector. Here we use the initial condition $\mathscr{B}(x, 0, \tau) = \mathscr{B}_{old}(x, \tau)$ in $\Omega \times [0, +\infty)$ and boundary condition $\mathscr{B}(x, t, 0) = \mathbb{I}$ in Q_T. Also, let us mention that for $k = 1$ we use backward Euler discretization instead of BDF-2, i.e. we set $\mathscr{B}_h^{n(-1)} = \underline{\underline{\boldsymbol{0}}}$, $\alpha_0 = 1$ and $\alpha_1 = -1$.

4 Numerical Experiment

We present two numerical experiments in the similar geometry but for different models. We consider a 2-dimensional problem of a fluid flow through a channel with an obstacle. The fluid enters the channel through the left vertical boundary and exits through the right one. At the inflow boundary we assume the following

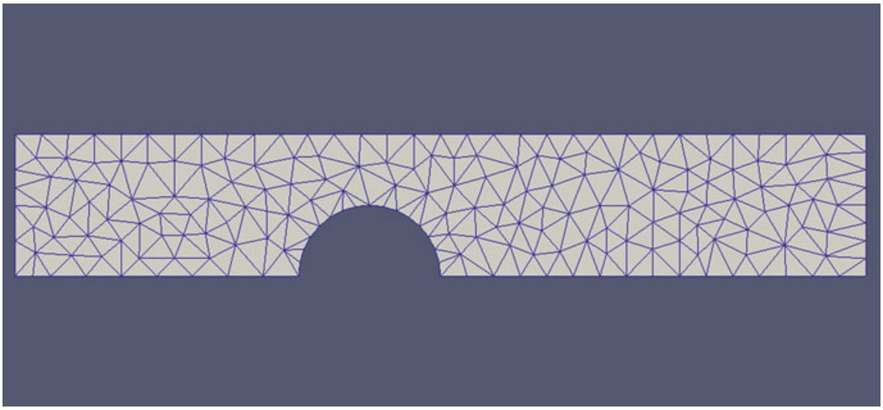


Fig. 1 The domain and its mesh for $h = 0.2$

Table 1 Half-step method error estimate for the UCM. The l parameter defines the level of discretization, i.e. $l = 1$ corresponds to the discretization parameters presented above, $l = 2$ corresponds to the discretization parameters of a half size, etc

l	$\|\boldsymbol{v}_l - \boldsymbol{v}_{l/2}\|$	$\|\pi_l - \pi_{l/2}\|$	$\|\mathscr{B}_l - B_{l/2}\|$
1.0	$1e-08$	$5e-11$	$1e-05$
0.5	$3e-09$	$5e-11$	$3e-06$
0.25	$8e-10$	$4e-11$	$8e-07$

parabolic velocity field

$$\boldsymbol{v}(x_1, x_2) = (x_2(1-x_2), 0).$$

On the outflow we assume a zero Neumann condition and on the rest of $\partial\Omega$ the velocity is set to $\mathbf{0}$. The domain with basic mesh is shown in Fig. 1.

Further, we set $h = 0.2$ and the time discretization steps to be equidistant with $\Delta t = 0.05, N = 8$. The choice of the age discretization steps loosely follows [2], i.e. $N_\tau = 43$ and $\Delta\tau_0 = 0.05$, $\Delta\tau_{k+1} = 1.0778\Delta\tau_k, k = 0, \dots, N_\tau - 1$. The polynomial degrees of approximations are set to $p = 2$ and $q = 1$.

In the first numerical experiment we test the method on the Upper Convected Maxwell (UCM) model, i.e. we assume that $\mu = 0, \mathscr{G}(\tau) = e^{-\tau}$ and $\mathscr{H}(\mathscr{B}) = \mathscr{B}$. Table 1 presents the error estimates obtained by the half-step method (i.e. in the error computation we take as a precise solution the solution of the problem with $h = h/2, \Delta t = \Delta t/2, N = 2N, N_\tau = 2N_\tau - 1$ and we insert new age nodes in the middle of each two following original nodes).

Table 2 Half-step method error estimate for Oldroyd-B

l	$\\|\boldsymbol{v}_l - \boldsymbol{v}_{l/2}\\|$	$\\|\pi_l - \pi_{l/2}\\|$	$\\|\mathscr{B}_l - B_{l/2}\\|$
1.0	$5e-11$	$4e-09$	$1e-05$
0.5	$6e-12$	$1e-09$	$3e-06$
0.25	$1e-12$	$3e-10$	$8e-07$

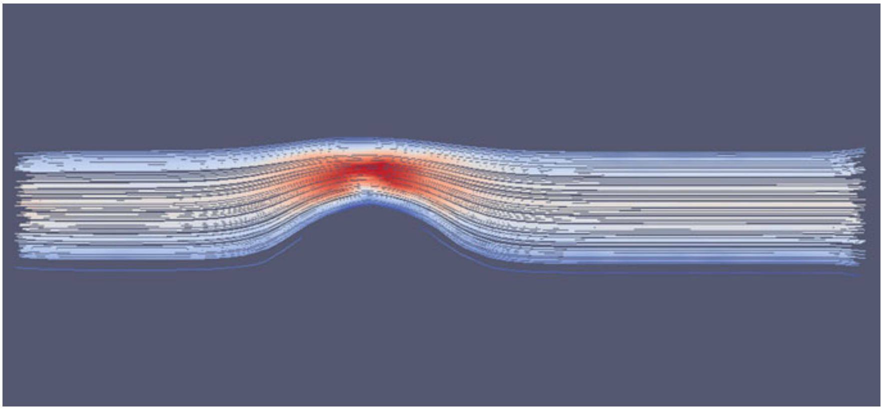


Fig. 2 The velocity field and streamlines for the first experiment. The *red colour* corresponds to higher velocity magnitude and the *blue colour* corresponds to lower velocity magnitude

The second numerical experiment is performed for the Oldroyd-B model with $\mu = 0.0035$ (i.e. the viscosity of blood), $\mathscr{G}(\tau) = e^{-\tau}$ and $\mathscr{H}(\mathscr{B}) = \mathscr{B}$. Table 2 presents the error estimates obtained by the half-step method.

5 Conclusion

Numerical experiments show that the method converges, measured by the half-step method. We also observe in Fig. 2 that the flow behaves in a reasonable way.

Thus, the method seems to be working although it was tested on simple problems. In future we plan to test the method on more benchmarks and more importantly to analyze the age discretization error and the spatial discretization error of the evolution equation of the Finger strain tensor.

Acknowledgements This work was supported by the Grant No. 454213 of the Grant Agency of the Charles University.

References

1. P.C. Bollada, T.N. Phillips, A modified deformation field method for integral constitutive models. J. Non-Newton. Fluid Mech. **163**, 78–87 (2009)
2. M.A. Hulsen, E.A.J.F. Peters, B.H.A.A. van den Brule, A new approach to the deformation fields method for solving complex flows using integral constitutive equations. J. Non-Newton. Fluid Mech. **98**, 201–221 (2001)
3. W.R. Hwang, M.A. Walkley, O.G. Harlen, A fast and efficient iterative scheme for viscoelastic flow simulations with the DEVSS finite element method. J. Non-Newton. Fluid Mech. **166**, 354–362 (2011)
4. E.A.J.F. Peters, M.A. Hulsen, B.H.A.A. van den Brule, Instationary Eulerian viscoelastic flow simulations using time separable Rivlin-Sawyers constitutive equations. J. Non-Newton. Fluid Mech. **89**, 209–228 (2000)
5. B. Riviére, *Discontinuous Galerkin Methods for Solving Elliptic and Parabolic Equations: Theory and Implementation*. Frontiers in Applied Mathematics (SIAM, Philadelphia, 2008)
6. A.P.G. van Heel, M.A. Hulsen, B.H.A.A. van den Brule, Simulation of the Doi-Edwards model in complex flow. J. Rheol. **43**, 1239–1260 (1999)

Stability Analysis of the ALE-STDGM for Linear Convection-Diffusion-Reaction Problems in Time-Dependent Domains

Monika Balázsová and Miloslav Feistauer

Abstract In this paper we investigate the stability of the space-time discontinuous Galerkin method (STDGM) for the solution of nonstationary, linear convection-diffusion-reaction problem in time-dependent domains formulated with the aid of the arbitrary Lagrangian-Eulerian (ALE) method. At first we define the continuous problem and reformulate it using the ALE method, which replaces the classical partial time derivative with the so called ALE-derivative and an additional convective term. In the second part of the paper we discretize our problem using the space-time discontinuous Galerkin method. The space discretization uses piecewise polynomial approximations of degree $p \geq 1$, in time we use only piecewise linear discretization. Finally in the third part of the paper we present our results concerning the unconditional stability of the method.

1 Formulation of the Continuous Problem

We consider an initial-boundary value nonstationary, linear convection-diffusion-reaction problem in a time-dependent bounded domain:

Find a function $u = u(x,t)$ with $x \in \Omega_t,\ t \in (0,T)$ such that

$$\frac{\partial u}{\partial t} + \boldsymbol{v} \cdot \nabla u - \epsilon \triangle u + cu = g \quad \text{in} \quad \Omega_t,\ t \in (0,T), \tag{1}$$

$$u = u_D \quad \text{on} \quad \partial\Omega_t,\ t \in (0,T), \tag{2}$$

$$u(x,0) = u^0(x), \quad x \in \Omega_0. \tag{3}$$

We assume that $\boldsymbol{v} = (v_1, v_2),\ c,\ g,\ u_D,\ u^0$ are given functions and $\epsilon > 0$ is a given constant. Moreover let $Q_T = \{(x,t);\ t \in (0,T),\ x \in \Omega_t\}$, and let us assume

M. Balázsová (✉) • M. Feistauer
Charles University in Prague, Sokolovská 83, 186 75 Praha 8, Czech Republic
e-mail: balazsova@karlin.mff.cuni.cz; feist@karlin.mff.cuni.cz

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_22

that there exist constants $c_v,\ c_c > 0$, such that

$$\begin{aligned} &\boldsymbol{v} \in C([0,T];\ W^{1,\infty}(\Omega_t)),\ |\nabla \boldsymbol{v}| \le c_v,\ |\boldsymbol{v}| \le c_v \quad \text{in} \quad Q_T,\\ &c \in C([0,T], L^{\infty}(\Omega_t)),\ |c(x,t)| \le c_c \quad \text{in} \quad Q_T. \end{aligned}$$

Problem (1)–(3) will be reformulated using the so called arbitrary Lagrangian-Eulerian (ALE) method. It is based on a regular one-to-one ALE mapping of the reference domain Ω_0 onto the current configuration Ω_t:

$$\begin{aligned} &\mathscr{A}_t : \overline{\Omega}_0 \to \overline{\Omega}_t,\\ &X \in \overline{\Omega}_0 \to x = x(X,t) = \mathscr{A}_t(X) \in \overline{\Omega}_t, \quad t \in [0,T]. \end{aligned}$$

We assume that $\mathscr{A}_t \in C^1([0,T]; W^{1,\infty}(\Omega_t))$, i.e. the mapping $\mathscr{A}_t$ belongs to the Bochner space of continuously differentiable functions in $[0,T]$ with values in the Sobolev space $W^{1,\infty}(\Omega_t)$. We define the ALE velocity by

$$\begin{aligned} &\tilde{\boldsymbol{z}}(X,t) = \frac{\partial}{\partial t}\mathscr{A}_t(X), \quad t \in [0,T],\ X \in \Omega_0,\\ &\boldsymbol{z}(x,t) = \tilde{\boldsymbol{z}}(\mathscr{A}_t^{-1}(x),t), \quad t \in [0,T],\ x \in \Omega_t. \end{aligned}$$

Let $|\boldsymbol{z}(x,t)|,\ |\mathrm{div}\, \boldsymbol{z}(x,t)| \le c_z$ for $x \in \Omega_t,\ t \in (0,T)$. Further, we define the ALE derivative $D_t f = Df/Dt$ of a function $f = f(x,t)$ for $x \in \Omega_t$ and $t \in [0,T]$ as

$$D_t f(x,t) = \frac{D}{Dt} f(x,t) = \frac{\partial \tilde{f}}{\partial t}(X,t),$$

where $\tilde{f}(X,t) = f(\mathscr{A}_t(X),t)$, $X \in \Omega_0$, and $x = \mathscr{A}_t(X) \in \Omega_t$. The use of the chain rule yields the relation

$$\frac{Df}{Dt} = \frac{\partial f}{\partial t} + \boldsymbol{z} \cdot \nabla f, \tag{4}$$

which allows us to reformulate problem (1)–(3) in the ALE form:

Find $u = u(x,t)$ with $x \in \Omega_t,\ t \in (0,T)$ such that

$$D_t u + (\boldsymbol{v} - \boldsymbol{z}) \cdot \nabla u - \epsilon \triangle u + cu = g \quad \text{in} \quad \Omega_t,\ t \in (0,T), \tag{5}$$

$$u = u_D \quad \text{on} \quad \partial\Omega_t, \tag{6}$$

$$u(x,0) = u^0(x), \quad x \in \Omega_0. \tag{7}$$

In what follows, we shall use the notation $\boldsymbol{w} = \boldsymbol{v} - \boldsymbol{z}$ for the ALE transport velocity.

Numerical methods for linear convection-diffusion-reaction equations in a domain Ω independent of time were analyzed e.g. in [5]. In the case, when problem (1)–(3) is considered in a fixed domain, error estimates for the space-time discontinuous Galerkin discretization were derived in [4]. These results were generalized to the case of nonlinear convection and diffusion (cf. [3]). The paper [1] is devoted to the proof of unconditional stability of the space-time discontinuous Galerkin method (STDGM) applied to nonlinear convection-diffusion problems. The STDGM was used with success for the numerical solution of compressible flow in time-dependent domains and also for the dynamical linear and nonlinear elasticity (see [3]). In [2], the stability of the time discontinuous Galerkin semidiscretization of problem (5)–(7) was analyzed. Here we are concerned with the investigation of the stability of the complete STDGM applied to problem (5)–(7) in a time-dependent domain.

2 Space-Time Semidiscretization

In the time interval $[0, T]$ we construct a partition formed by time instants $0 = t_0 < t_1 < \ldots < t_M = T$ and set $I_m = (t_{m-1}, t_m)$ and $\tau_m = t_m - t_{m-1}$ for $m = 1, \ldots, M$. Further we set $\tau = \max_{m=1,\cdots,M} \tau_m$. For a function φ defined in $\bigcup_{m=1}^{M} I_m$ we denote one-sided limits at t_m as $\varphi_m^{\pm} = \varphi(t_m\pm) = \lim_{t\to t_m\pm} \varphi(t)$ and the jump as $\{\varphi\}_m = \varphi(t_m+) - \varphi(t_m-)$.

For any $t \in [0, T]$ we denote by $\mathscr{T}_{h,t}$ a partition of the closure $\overline{\Omega}_t$ into a finite number of closed triangles with mutually disjoint interiors. We set $h_K = \mathrm{diam}(K)$ for $K \in \mathscr{T}_{h,t}$. The boundary of the domain will be divided into two parts: $\partial\Omega_t = \partial\Omega_t^- \cup \partial\Omega_t^+$:

$$\boldsymbol{w}(x, t) \cdot \boldsymbol{n}(x) < 0 \text{ on } \partial\Omega_t^-, \forall t \in [0, T] \text{ (inflow boundary)}$$
$$\boldsymbol{w}(x, t) \cdot \boldsymbol{n}(x) \geq 0 \text{ on } \partial\Omega_t^+, \forall t \in [0, T] \text{ (outflow boundary)},$$

where $\boldsymbol{n}$ denotes the unit outer normal to ∂K. Similarly for each $K \in \mathscr{T}_{h,t}$ we set

$$\partial K^-(t) = \{x \in \partial K;\ \boldsymbol{w}(x, t) \cdot \boldsymbol{n}(x) < 0\},$$
$$\partial K^+(t) = \{x \in \partial K;\ \boldsymbol{w}(x, t) \cdot \boldsymbol{n}(x) \geq 0\}.$$

By $\mathscr{F}_{h,t}$ we denote the system of all faces of all elements $K \in \mathscr{T}_{h,t}$. It consists of the set of all inner faces $\mathscr{F}_{h,t}^I$ and the set of all boundary faces $\mathscr{F}_{h,t}^B$: $\mathscr{F}_{h,t} = \mathscr{F}_{h,t}^I \cup \mathscr{F}_{h,t}^B$. Each $\Gamma \in \mathscr{F}_{h,t}$ will be associated with a unit normal vector $\boldsymbol{n}_\Gamma$. By $K_\Gamma^{(L)}$ and $K_\Gamma^{(R)} \in \mathscr{T}_{h,t}$ we denote the elements adjacent to the face $\Gamma \in \mathscr{F}_{h,t}$. We shall use the convention that $\boldsymbol{n}_\Gamma$ is the outer normal to $\partial K_\Gamma^{(L)}$. Over a triangulation $\mathscr{T}_{h,t}$, for each positive integer k, we define the broken Sobolev space $H^k(\Omega_t, \mathscr{T}_{h,t}) = \{\varphi; \varphi|_K \in H^k(K) \quad \forall K \in \mathscr{T}_{h,t}\}$.

If $\varphi \in H^1(\Omega_t, \mathscr{T}_{h,t})$ and $\Gamma \in \mathscr{F}_{h,t}$, then $\varphi|_\Gamma^{(L)}, \varphi|_\Gamma^{(R)}$ will denote the traces of φ on Γ from the side of elements $K_\Gamma^{(L)}, K_\Gamma^{(R)}$ adjacent to Γ. For $\Gamma \in \mathscr{F}_{h,t}^I$ we set

$$\langle \varphi \rangle_\Gamma = \frac{1}{2}\left(\varphi|_\Gamma^{(L)} + \varphi|_\Gamma^{(R)}\right), \quad [\varphi]_\Gamma = \varphi|_\Gamma^{(L)} - \varphi|_\Gamma^{(R)},$$

$$h(\Gamma) = \frac{h_{K_\Gamma^{(L)}} + h_{K_\Gamma^{(R)}}}{2} \quad \text{for } \Gamma \in \mathscr{F}_{h,t}^I, \quad h(\Gamma) = h_{K_\Gamma^{(L)}} \quad \text{for } \Gamma \in \mathscr{F}_{h,t}^B.$$

If $u, \varphi \in H^2(\Omega_t, \mathscr{T}_{h,t})$, $\theta \in \mathbb{R}$ and $c_W > 0$, we introduce the following forms.

Convection form:

$$b_h(u,\varphi,t) = \sum_{K \in \mathscr{T}_{h,t}} \int_K \boldsymbol{w} \cdot \nabla u \, \varphi \, dx$$
$$- \sum_{K \in \mathscr{T}_{h,t}} \int_{\partial K^- \cap \partial \Omega_t} \boldsymbol{w} \cdot \boldsymbol{n} u \varphi \, dS - \sum_{K \in \mathscr{T}_{h,t}} \int_{\partial K^- \setminus \partial \Omega_t} \boldsymbol{w} \cdot \boldsymbol{n} [u] \varphi \, dS,$$

Diffusion form:

$$a_h(u,\varphi,t) = \sum_{K \in \mathscr{T}_{h,t}} \int_K \nabla u \cdot \nabla \varphi \, dx$$
$$- \sum_{\Gamma \in \mathscr{F}_{h,t}^I} \int_\Gamma \left(\langle \nabla u \rangle \cdot \boldsymbol{n}_\Gamma \, [\varphi] + \theta \, \langle \nabla \varphi \rangle \cdot \boldsymbol{n}_\Gamma \, [u] \right) \, dS$$
$$- \sum_{K \in \mathscr{T}_{h,t}} \int_{\partial K^- \cap \partial \Omega_t} \left(\nabla u \cdot \boldsymbol{n}_\Gamma \, \varphi + \theta \nabla \varphi \cdot \boldsymbol{n}_\Gamma \, u - \theta \nabla \varphi \cdot \boldsymbol{n}_\Gamma \, u_D \right) \, dS,$$

Interior and boundary penalty:

$$J_h(u,\varphi,t) = c_W \sum_{\Gamma \in \mathscr{F}_{h,t}^I} h(\Gamma)^{-1} \int_\Gamma [u] \, [\varphi] \, dS$$
$$+ c_W \sum_{K \in \mathscr{T}_{h,t}} h(\Gamma)^{-1} \int_{\partial K^- \cap \partial \Omega_t} u \, \varphi \, dS,$$

$$A_h(u,\varphi,t) = \epsilon a_h(u,\varphi,t) + \epsilon \, J_h(u,\varphi,t),$$

Reaction form:

$$c_h(u,\varphi,t) = \sum_{K \in \mathscr{T}_{h,t}} \int_K c u \varphi \, dx,$$

Right-hand side form:

$$l_h(\varphi, t) = \sum_{K \in \mathscr{T}_{h,t}} \int_K g\varphi \, dx + \epsilon \, c_W \sum_{\Gamma \in \mathscr{F}_{h,t}^B} h(\Gamma)^{-1} \int_\Gamma u_D \, \varphi \, dS.$$

Let us note that in integrals over faces we omit the subscript Γ. We consider $\theta = 1$, $\theta = 0$ and $\theta = -1$ and get the symmetric (SIPG), incomplete (IIPG) and nonsymmetric (NIPG) variants of the approximation of the diffusion terms, respectively.

Further, we set

$$(\varphi, \psi)_\omega = \int_\omega \varphi \psi \, dx, \quad \|\varphi\|_\omega = \left(\int_\omega |\varphi|^2 \, dx \right)^{1/2},$$
$$\|\eta\|_{\boldsymbol{w},\sigma} = \left\| \sqrt{|\boldsymbol{w} \cdot \boldsymbol{n}|} \, \eta \right\|_{L^2(\sigma)},$$

where $\omega \subset \mathbb{R}^2$, σ is either a subset of $\partial\Omega$ or ∂K and $\boldsymbol{n}$ denotes the corresponding outer unit normal to $\partial\Omega$ or ∂K, provided the integrals make sense.

Let $p, \ q \geq 1$ be integers. For any $m = 1, \dots, M$ and $t \in [0, T]$ we define the finite-dimensional spaces

$$S_{h,t}^p = \left\{ \varphi \in L^2(\Omega_t); \ \varphi|_K \in P^p(K), \ K \in \mathscr{T}_{h,t}, \ t \in [0, T] \right\},$$
$$S_{h,\tau}^{p,q} = \Big\{ \varphi \in L^2(Q_T); \ \varphi = \varphi(x, t), \ \text{for each } X \in \Omega_0$$
$$\text{the function } \varphi(\mathscr{A}_t(X), t) \text{ is a polynomial}$$
$$\text{of degree } \leq q \text{ in } t, \ \varphi(\cdot, t) \in S_{h,t}^p \text{ for every } t \in I_m, \ m = 1, \dots, M \Big\}.$$

Definition 1 We say that function U is an approximate solution of problem (5)–(7), if $U \in S_{h,\tau}^{p,q}$ and

$$\int_{I_m} \left((D_t U, \varphi)_{\Omega_t} + A_h(U, \varphi, t) + b_h(U, \varphi, t) + c_h(U, \varphi, t) \right) dt \tag{8}$$
$$+ (\{U\}_{m-1}, \varphi_{m-1}^+)_{\Omega_{t_{m-1}}} = \int_{I_m} l_h(\varphi, t) \, dt \quad \forall \varphi \in S_{h,\tau}^{p,q}, \quad m = 1, \dots, M,$$
$$U_0^- \in S_{h,0}^p, \quad (U_0^- - u^0, v_h) = 0 \quad \forall v_h \in S_{h,0}^p. \tag{9}$$

3 Analysis of the Stability

In our further considerations for each $t \in [0, T]$ we introduce a system of conforming triangulations $\{\mathscr{T}_{h,t}\}_{h \in (0, h_0)}$, where $h_0 > 0$. We assume that it is shape regular and locally quasiuniform. Under these assumptions, the multiplicative trace inequality and the inverse inequality hold.

Moreover, we assume that $\mathscr{T}_{h,t} = \{K_t = \mathscr{A}_t(K_0); K_0 \in \mathscr{T}_{h,0}\}$. This assumption is usually satisfied in practical computations, when the ALE mapping $\mathscr{A}_t$ is a continuous, piecewise affine mapping in $\overline{\Omega}_0$ for each $t \in [0, T]$.

In the space $H^1(\Omega, \mathscr{T}_{h,t})$ we define the norm

$$\|\varphi\|_{DG,t} = \left(\sum_{K \in \mathscr{T}_{h,t}} |\varphi|^2_{H^1(K)} + J_h(\varphi, \varphi, t) \right)^{1/2}.$$

Moreover, over $\partial\Omega$ we define the norm

$$\|u_D\|_{DGB,t} = \left(c_W \sum_{K \in \mathscr{T}_{h,t}} h^{-1}(\Gamma) \int_{\partial K^- \cap \partial\Omega_t} |u_D|^2 \, dS \right)^{1/2}.$$

If we use $\varphi := U$ as a test function in (8), we get the basic identity

$$\int_{I_m} \big((D_t U, U)_{\Omega_t} + A_h(U, U, t) + b_h(U, U, t) + c_h(U, U, t) \big) \, dt \tag{10}$$
$$+ (\{U\}_{m-1}, U^+_{m-1})_{\Omega_{t_{m-1}}} = \int_{I_m} l_h(U, t) \, dt.$$

Let us denote

$$\sigma(U) = \frac{1}{2} \sum_{K \in \mathscr{T}_{h,t}} \left(\|U\|^2_{\boldsymbol{w}, \partial K \cap \partial\Omega} + \|[U]\|^2_{\boldsymbol{w}, \partial K^- \setminus \partial\Omega} \right). \tag{11}$$

For a sufficiently large constant c_W, whose lower bound is determined by the constants from the multiplicative trace inequality, inverse inequality and local quasiuniformity of the meshes, we can prove the coercivity of the diffusion and penalty terms:

$$\int_{I_m} A_h(U, U, t) \, dt \geq \frac{\epsilon}{2} \int_{I_m} \|U\|^2_{DG,t} \, dt - \frac{\epsilon}{2} \int_{I_m} \|u_D\|^2_{DGB,t} \, dt. \tag{12}$$

Furthermore, if $k_1 > 0$, then the following inequalities for the convective term, reaction term and for the right-hand side form hold:

$$b_h(U, U, t) = \sigma(U) - \frac{1}{2} \int_{\Omega_t} U^2 \nabla \cdot \boldsymbol{w} \, dx, \tag{13}$$

$$\int_{I_m} |c_h(U, U, t)| \, dt \leq c_c \int_{I_m} \|U\|^2_{\Omega_t} \, dt, \tag{14}$$

$$\int_{I_m} |l_h(U,t)|\, dt \le \frac{1}{2}\int_{I_m} \left(\|g\|^2_{\Omega_t} + \|U\|^2_{\Omega_t}\right)\, dt \tag{15}$$
$$+\epsilon k_1 \int_{I_m} \|u_D\|^2_{DGB,t}\, dt + \frac{\epsilon}{k_1}\int_{I_m} \|U\|^2_{DG,t}\, dt.$$

In what follows, we are concerned with the derivation of inequalities based on estimating the expression $\int_{I_m}(D_t U, U)_{\Omega_t}\, dt$. By some manipulation we find that

$$\int_{I_m} (D_t U, U)_{\Omega_t}\, dt + \left(\{U\}_{m-1}, U^+_{m-1}\right)_{\Omega_{t_{m-1}}} \tag{16}$$
$$\ge \frac{1}{2}\left(\|U^-_m\|^2_{\Omega_{t_m}} - \|U^-_{m-1}\|^2_{\Omega_{t_{m-1}}} + \|\{U\}_{m-1}\|^2_{\Omega_{t_{m-1}}}\right)$$
$$-\frac{1}{2}\int_{I_m} (U^2, \nabla\cdot \boldsymbol{z})_{\Omega_t}\, dt,$$

and

$$\int_{I_m} (D_t U, U)_{\Omega_t}\, dt + (\{U\}_{m-1}, U^+_{m-1})_{\Omega_{t_{m-1}}} \tag{17}$$
$$\ge \frac{1}{2}(\|U^-_m\|^2_{\Omega_{t_m}} + \frac{1}{2}\|U^+_{m-1}\|^2_{\Omega_{t_{m-1}}}) - (U^-_{m-1}, U^+_{m-1})_{\Omega_{t_{m-1}}}$$
$$-\frac{1}{2}\int_{I_m} (U^2, \nabla\cdot \boldsymbol{z})_{\Omega_t}\, dt.$$

Taking into account that $\sigma(U) \ge 0$ and $\boldsymbol{w} = \boldsymbol{v} - \boldsymbol{z}$, from (10), (14) and (12)–(16) and putting $k_1 = 4$, we get the relation

$$\|U^-_m\|^2_{\Omega_{t_m}} - \|U^-_{m-1}\|^2_{\Omega_{t_{m-1}}} - \int_{I_m} (U^2, \nabla\cdot \boldsymbol{v})_{\Omega_t}\, dt \tag{18}$$
$$+\int_{I_m} (2c-1, U^2)_{\Omega_t} + \frac{\epsilon}{2}\int_{I_m} \|U\|^2_{DG,t}\, dt$$
$$\le c_1 \int_{I_m} \left(\|g\|^2_{\Omega_t} + \|u_D\|^2_{DGB,t}\right)\, dt$$

with a constant c_1 independent of data, h and τ.

First, let us assume that

$$2c - \nabla\cdot \boldsymbol{v} \ge 1. \tag{19}$$

Then the summation of (18) over $m = 1, \dots, k \ \leq M$ yields the estimate

$$\|U_k^-\|_{\Omega_{t_k}} + \frac{\epsilon}{2} \sum_{m-1}^{k} \int_{I_m} \|U\|_{DG,t}^2 \, dt \tag{20}$$

$$\leq \|U_0^-\|_{\Omega_0}^2 + c_1 \sum_{m-1}^{k} \int_{I_m} \left(\|g\|_{\Omega_t}^2 + \|u_D\|_{DGB,t}^2 \right) dt,$$

which proves the stability.

If condition (19) is not valid, then the stability analysis is more complicated. In this case, instead of (18) we get the inequality

$$\|U_m^-\|_{\Omega_{t_m}}^2 - \|U_{m-1}^-\|_{\Omega_{t_{m-1}}}^2 + \frac{\epsilon}{2} \int_{I_m} \|U\|_{DG,t}^2 \, dt \tag{21}$$

$$\leq c_1 \sum_{m-1}^{k} \int_{I_m} \left(\|g\|_{\Omega_t}^2 + \|u_D\|_{DGB,t}^2 \right) dt + c_2 \int_{I_m} \|U\|_{\Omega_t}^2 \, dt.$$

It is necessary to estimate the term $\int_{I_m} \|U\|_{\Omega_t}^2 \, dt$. It is rather technical and the proof has been carried out for $q = 1$, i.e., for piecewise linear time discretization. Then it is possible to show that there exist constants L_1 and M_1 such that

$$\|U_{m-1}^+\|_{\Omega_{t_{m-1}}}^2 + \|U_m^-\|_{\Omega_{t_m}}^2 \geq \frac{L_1}{\tau_m} \int_{I_m} \|U\|_{\Omega_t}^2 \, dt, \tag{22}$$

$$\|U_{m-1}^+\|_{\Omega_{t_{m-1}}}^2 \leq \frac{M_1}{\tau_m} \int_{I_m} \|U\|_{\Omega_t}^2 \, dt.$$

This allows to prove that there exists a constant $c^* > 0$ depending on c_2 and L_1 such that

$$\int_{I_m} \|U\|_{\Omega_t}^2 dt \leq \frac{2c_1}{L_1} \tau_m \int_{I_m} \left(\|g\|_{\Omega_t}^2 + \|u_D\|_{DGB,t}^2 \right) dt + \frac{8M_1}{L_1^2} \tau_m \|U_{m-1}^-\|_{\Omega_{t_{m-1}}}^2 \tag{23}$$

holds, if $0 < \tau_m \leq c^*$.

Now, by virtue of (21) and (23), the summation over $m = 1, \dots, k \ \leq M$ and the application of the discrete Gronwall lemma we get the following result.

Theorem 2 *Let* $q = 1$ *and* $0 < \tau_m \le c^*$. *Then there exists a constant* $c_3 > 0$ *such that*

$$\|U_m^-\|^2_{\Omega_{t_m}} + \sum_{j=1}^{m} \|\{U_{j-1}\}\|^2_{\Omega_{t_{j-1}}} + \frac{\beta_0}{2}\sum_{j=1}^{m} \int_{I_j} \|U\|^2_{DG,j}\,dt \tag{24}$$

$$\le c_3 \left(\|U_0^-\|^2_{\Omega_{t_0}} + \sum_{j=1}^{m} \int_{I_j} R_j\,dt \right), \ m = 1, \dots, M, \ h \in (0, h_0),$$

where

$$R_j = c_1 \left(1 + \frac{2c_2}{L_1}\tau_j\right) \left(\|g\|^2_{\Omega_j} + \|u_D\|^2_{DGB,t} \right).$$

Acknowledgements This research was supported by the grant 13-00522S of the Czech Science Foundation (M. Feistauer) and by the project GA UK No. 127615 of the Charles University in Prague (M. Balázsová).

References

1. M. Balázsová, M. Feistauer, M. Hadrava, A. Kosík, On the stability of the space-time discontinuous Galerkin method for the numerical solution of nonstationary nonlinear convection-diffusion problems. J. Numer. Math. **23**(3), 211–233 (2015)
2. A. Bonito, I. Kyza, R.H. Nochetto, Time-discrete higher-order ALE formulations: stability. SIAM J. Numer. Anal. **51**(1), 577–604 (2013)
3. V. Dolejší, M. Feistauer, *Discontinuous Galerkin Method – Analysis and Applications to Compressible Flow* (Springer, Cham, 2015)
4. M. Feistauer, J. Hájek, K. Švadlenka, Space-time discontinuous Galerkin method for solving nonstationary linear convection-diffusion-reaction problems. Appl. Math. **52**, 197–233 (2007)
5. H.-G. Roos, M. Stynes, L. Tobiska, *Robust Numerical Methods for Singularly Perturbed Differential Equations* (Springer, Berlin, 2008)

A Posteriori Error Estimates for Nonstationary Problems

Vít Dolejší, Filip Roskovec, and Miloslav Vlasák

Abstract We apply continuous and discontinuous Galerkin time discretization together with standard finite element method for space discretization to the heat equation. For the numerical solution arising from these discretizations we present a guaranteed and fully computable a posteriori error upper bound. Moreover, we present local asymptotic efficiency estimate of this bound.

1 Introduction

We consider the heat equation, which represents a model problem to more general linear parabolic problems. We discretize this problem by standard finite element method in space and by either continuous or discontinuous Galerkin method in time.

Recently, time discretizations of Galerkin type start to be very popular. They represent higher order and very robust schemes for solving ordinary differential equations. When combined with classical Galerkin space discretizations, e.g. with finite element method (FEM), it is possible to analyze the complete discretization in a unified framework. For a survey about Galerkin time discretizations see [1] and [2]. A nice result presenting the connection of these discretizations to classical Runge–Kutta methods can be found in [8].

In this paper we shall focus on a posteriori error analysis of proposed problem. Our aim is to present a guaranteed, cheap and fully computable upper bound to chosen error measure that provides local efficiency at least asymptotically. To achieve these properties we use the technique of so-called equilibrated flux reconstruction, see e.g. [5]. We have been influenced by [4], where lower order time discretizations are considered, and by [2], where Galerkin time discretizations are analyzed and nodal superconvergence is derived via a posteriori error estimates.

V. Dolejší • F. Roskovec • M. Vlasák (✉)
Faculty of Mathematics and Physics, Charles University in Prague, Sokolovska 83, 186 75 Prague 8, Czech Republic
e-mail: dolejsi@karlin.mff.cuni.cz; roskovec@gmail.com; vlasak@karlin.mff.cuni.cz

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_23

2 Continuous Problem

Let $\Omega \subset R^d$ $(d = 1, 2, 3)$ be a bounded polyhedral domain with Lipschitz continuous boundary $\partial\Omega$ and $T > 0$. Let us consider the following initial–boundary value problem

$$\begin{aligned} \frac{\partial u}{\partial t} - \Delta u &= f \quad \text{in } \Omega \times (0, T), \\ u &= 0 \quad \text{in } \partial\Omega \times (0, T), \\ u &= u^0 \quad \text{in } \Omega. \end{aligned} \tag{1}$$

We assume that the right-hand side $f \in C(0, T, L^2(\Omega))$ and the initial condition $u^0 \in L^2(\Omega)$.

Let $(.,.)$ and $\|.\|$ be the $L^2(\Omega)$-scalar product and norm, respectively. Let us denote the time derivative $u' = \frac{\partial u}{\partial t}$. We define spaces $X = L^2(0, T, H_0^1(\Omega))$ and

$$\begin{aligned} Y &= \{v \in X : v' \in L^2(0, T, L^2(\Omega))\}, \\ Y_0 &= \{v \in X : v' \in L^2(0, T, L^2(\Omega)), v(0) = u^0\}. \end{aligned} \tag{2}$$

It is well known that the spaces Y and Y_0 are subsets of $C([0, T], L^2(\Omega))$.

Definition 1 We call $u \in Y_0$ the weak solution of problem (1), if

$$\int_0^T (f, v) - (u', v) - (\nabla u, \nabla v) dt = 0, \quad \forall v \in X. \tag{3}$$

We assume that there exists a unique weak solution of problem (3).

3 Discretization

We consider a space partition $\mathscr{T}_h$ consisting of a finite number of closed, d - dimensional simplices K with mutually disjoint interiors and covering $\overline{\Omega}$, i.e. $\overline{\Omega} = \cup_{K \in \mathscr{T}_h} K$. We assume conforming properties, i.e. neighbouring elements share an entire edge or face. We set $h_K = \text{diam}(K)$ and $h = \max_K h_K$. By ρ_K we denote the radius of the largest d-dimensional ball inscribed into K. We assume shape regularity of elements, i.e. $h_K/\rho_K \leq C$ for all $K \in \mathscr{T}_h$, where the constant does not depend on $\mathscr{T}_h$ for $h \in (0, h_0)$.

We set the space for the semidiscrete solution

$$X_h = \{v \in H_0^1(\Omega) : v|_K \in P^p(K)\}, \tag{4}$$

where $P^p(K)$ denotes the space of polynomials up to the degree $p \geq 1$ on K. We define $\Pi_\Omega^p : H_0^1(\Omega) \to X_h$ to be the L^2-orthogonal projection.

In order to discretize problem (3) in time, we consider a time partition $0 = t_0 < t_1 < \ldots < t_r = T$ with time intervals $I_m = (t_{m-1}, t_m)$, time steps $\tau_m = |I_m| = t_m - t_{m-1}$ and $\tau = \max_{m=1,\ldots,r} \tau_m$. Let $(.,.)_{K,m}$ and $(.,.)_K$ be the local L^2-scalar products over $K \times I_m$ and K, respectively, and $\|.\|_{K,m}$ be the local $L^2(K \times I_m)$-norm. In the forthcoming discretization process we will assume two variants of the time discretization, the conforming and the nonconforming one. In the conforming case, the approximate solution will be sought in the spaces of piecewise polynomial functions

$$Y_{0h}^\tau = \{v \in Y : v|_{I_m} = \sum_{j=0}^{q+1} v_{j,m} t^j,\ v_{j,m} \in X_h, v(0) = \Pi_\Omega^p u^0\} \tag{5}$$

and in the nonconforming case in the space

$$X_h^\tau = \{v \in X : v|_{I_m} = \sum_{j=0}^{q} v_{j,m} t^j,\ v_{j,m} \in X_h\}. \tag{6}$$

The spaces Y_{0h}^τ and X_h^τ represent natural discrete spaces to Y_0 and X, respectively. The space Y_{0h}^τ consists of functions that are one degree higher in time than the functions from the space X_h^τ. On the other hand the functions from Y_{0h}^τ are continuous with respect to time with fixed starting value at 0. Altogether, both these spaces have the same dimension $r(q+1)\dim X_h$.

For a function $v \in X_h^\tau$ we define the one–sided limits

$$v_\pm^m = v(t_m \pm) = \lim_{t \to t_m \pm} v(t) \tag{7}$$

and the jumps

$$\{v\}_m = v_+^m - v_-^m, \quad m \geq 1 \quad \text{and} \quad \{v\}_0 = v_+^0 - u^0. \tag{8}$$

We omit the subscript ± for continuous functions $v \in Y$, since $v(t_m \pm) = v(t_m)$.

Now, we are able to formulate two variants of discrete schemes – the conforming version:

Definition 2 We say that the function $u_h^\tau \in Y_{0h}^\tau$ is the discrete solution of problem (3) obtained by time continuous Galerkin – finite element method (cG–FEM), if the following conditions are satisfied

$$\int_{I_m} ((u_h^\tau)', v) + (\nabla u_h^\tau, \nabla v) dt = \int_{I_m} (f, v) dt \tag{9}$$

$$\forall m = 1, \ldots, r,\ \forall v \in X_h^\tau,$$

and the nonconforming version:

Definition 3 We say that the function $u_h^\tau \in X_h^\tau$ is the discrete solution of problem (3) obtained by time discontinuous Galerkin – finite element method (dG–FEM), if the following conditions are satisfied

$$\int_{I_m} ((u_h^\tau)', v) + (\nabla u_h^\tau, \nabla v) dt + (\{u_h^\tau\}_{m-1}, v_+^{m-1}) = \int_{I_m} (f, v) dt \qquad (10)$$
$$\forall m = 1, \ldots, r,\ \forall v \in X_h^\tau.$$

It is evident that the exact solution $u \in Y_0$ defined by (3) satisfies both relations (9) and (10).

The methods (9) and (10) can be viewed as a generalization of classical one–step methods for parabolic problems. It is possible to show that setting $q = 0$, i.e. piecewise linear continuous approximation in time for cG–FEM or piecewise constant approximation in time for dG–FEM, is equivalent (up to suitable quadrature of the right–hand side) to Crank–Nicolson, resp. backward Euler method, in time and FEM in space.

4 A Posteriori Error Analysis

In this section we shall propose suitable error measure and derive a posteriori error estimate of this measure.

4.1 Error Measure

Let $d_{K,m} > 0$ be an arbitrary parameter associated with space-time element $K \times I_m$, e.g. $d_{K,m}^2 = h_K^2 + \tau_m^2$ or $d_{K,m} = 1$ or $d_{K,m} = h_K$ or $d_{K,m}^2 = (h_K^{-2} + \tau_m^{-2})^{-1}$. Let us define the space

$$Y^\tau = \{v \in X : v'|_{I_m} \in L^2(I_m, L^2(\Omega))\} \qquad (11)$$

of piecewise continuous functions with respect to time. We define the norm

$$\|v\|_{Z,K,m}^2 = \frac{h_K^2 \|\nabla v\|_{K,m}^2 + \tau_m^2 \|v'\|_{K,m}^2}{d_{K,m}^2}, \quad \|v\|_Z^2 = \sum_{K,m} \|v\|_{Z,K,m}^2. \qquad (12)$$

Since $Y^\tau \subset X$, we gain from (3) that the exact solution $u \in Y$ satisfies

$$\int_{I_m} (f, v) - (u', v) - (\nabla u, \nabla v) dt - (\{u\}_{m-1}, v_+^{m-1}) = 0 \tag{13}$$
$$\forall m = 1, \ldots, r, \ \forall v \in Y^\tau.$$

The existence of the solution u of problem (13) comes clearly from the existence of the solution of problem (3). We shall focus on uniqueness of the solution of problem (13). Let us assume that there exists another solution $u_1 \in Y^\tau$ of problem (13). After subtracting the equation for u from the equation for u_1 and setting $v = 2(u - u_1)$ we gain

$$\begin{aligned} 0 = &\int_{I_m} 2(u' - u_1', u - u_1) + 2\|\nabla(u - u_1)\|^2 dt \\ &+2(\{u - u_1\}_{m-1}, (u - u_1)_+^{m-1}) \\ = &\ \|(u - u_1)_-^m\|^2 - \|(u - u_1)_-^{m-1}\|^2 + \|\{u - u_1\}_{m-1}\|^2 \\ &+2\int_{I_m} \|\nabla(u - u_1)\|^2 dt \end{aligned} \tag{14}$$

Summing this relation over $m = 1, \ldots, r$ and using the fact $u_-^0 = u^0 = u_{1-}^0$ we gain

$$\|(u - u_1)_-^r\|^2 + \sum_{m=1}^r \|\{u - u_1\}_{m-1}\|^2 + 2\int_0^T \|\nabla(u - u_1)\|^2 dt = 0, \tag{15}$$

which implies $u = u_1$.

It is natural to define error measure EST for both variants of discretization as residual of (13)

$$\begin{aligned} \mathrm{EST}(w) = \sup_{0 \neq v \in Y^\tau} \frac{1}{\|v\|_Z} \Bigg(&\sum_{K,m} (f, v)_{K,m} - (w', v)_{K,m} \\ &-(\nabla w, \nabla v)_{K,m} - (\{w\}_{m-1}, v_+^{m-1})_K \Bigg) \end{aligned} \tag{16}$$

for $w \in X_h^\tau$.

It is possible to show that the uniqueness of the solution of problem (13) implies that $\mathrm{EST}(u_h^\tau) = 0$, if and only if u_h^τ is equal to the exact solution u.

4.2 Reconstruction of the Solution with Respect to Time

Since the exact solution $u \in Y_0 \subset C([0,T], L^2(\Omega))$, i.e. u is continuous in time and $u(0) = u^0$, we will reconstruct the discrete solution u_h^τ in such a way, that the reconstruction satisfies these properties too. For conforming variant of discretization (cG-FEM) this task is easier, since the solution is already continuous in time, but the initial condition can still be violated. For nonconforming version we need to reconstruct for both reasons.

Let $r_m \in P^{q+1}(I_m)$ be the right Radau polynomial on I_m, i.e. $r_m(t_{m-1}) = 1$, $r_m(t_m) = 0$ and r_m is orthogonal to P^{q-1}. Then there exists a polynomial reconstruction $R_h^\tau = R_h^\tau(u_h^\tau)$ for both variants of discretization such that

$$R_h^\tau(t) = u_h^\tau(t) - \{u_h^\tau\}_{m-1} r_m(t), \quad \forall t \in I_m. \tag{17}$$

Since the cG-FEM solution u_h^τ is continuous in time, the reconstruction R_h^τ is equal to u_h^τ except I_1. It is still necessary to reconstruct the discrete initial condition $\Pi_\Omega^p u^0$ on I_1, see (8).

The resulting function R_h^τ is continuous in time and satisfies the initial condition, i.e. $R_h^\tau \in Y_0$. Moreover,

$$\begin{aligned} \int_{I_m} ((R_h^\tau)', v)dt &= \int_{I_m} ((u_h^\tau)', v) - r_m'(\{u_h^\tau\}_{m-1}, v)dt \\ &= \int_{I_m} ((u_h^\tau)', v)dt + \int_{I_m} r_m(\{u_h^\tau\}_{m-1}, v')dt \\ &\quad -r_m(t_m)(\{u_h^\tau\}_{m-1}, v_-^m) + r_m(t_{m-1})(\{u_h^\tau\}_{m-1}, v_+^{m-1}) \\ &= \int_{I_m} ((u_h^\tau)', v)dt + (\{u_h^\tau\}_{m-1}, v_+^{m-1}), \quad \forall v \in P^q(I_m, L^2(\Omega)). \end{aligned} \tag{18}$$

Such a reconstruction is used to show the equivalence among Radau IIA Runge–Kutta method, Radau collocation method and discontinuous Galerkin method. For the details see, e.g. [6] and [7]. Such a reconstruction is also used for proving a posteriori nodal superconvergence in [2].

4.3 Reconstruction of the Solution with Respect to Space

It is possible to show that the exact solution satisfies $\nabla u \in L^2(0,T,H(\text{div}))$. Since $\nabla u_h^\tau \notin L^2(0,T,H(\text{div}))$ in general, we reconstruct also ∇u_h^τ. Let $RTN_p(K)$ be the Raviar-Thomas-Nedelec space of order p, i.e. $RTN_p(K) = P_p(K)^d + xP_p(K)$. Let us denote the patch $\mathscr{T}_a = \bigcup_{a\in K} K$ of vertex a. Then we can define *RTN* spaces on $\mathscr{T}_a$

$$RTN_p^{N,0}(\mathscr{T}_a) = \{v \in RTN_p(\mathscr{T}_a) : v \cdot n = 0 \ \forall e \subset \partial\mathscr{T}_a\}, \quad a \notin \partial\Omega, \tag{19}$$
$$RTN_p^{N,0}(\mathscr{T}_a) = \{v \in RTN_p(\mathscr{T}_a) : v \cdot n = 0 \ \forall e \subset \partial\mathscr{T}_a \setminus \partial\Omega\}, \quad a \in \partial\Omega.$$

Let us denote by $P_*^p(\mathscr{T}_a)$ piecewise polynomials of order p for $a \in \partial\Omega$. Moreover, $P_*^p(\mathscr{T}_a)$ consists of functions with zero mean value for $a \notin \partial\Omega$. Let us denote ψ_a piecewise linear "hat" function associated with vertex a with $\psi(a) = 1$, $\psi = 0$ on $\partial\mathscr{T}_a$.

We formulate space–time version of patch–wise reconstruction from [3]. We seek $\sigma_a^\tau|_{\mathscr{T}_a \times I_m} \in P^q(I_m, RTN_p^{N,0}(\mathscr{T}_a))$ and $r_a^\tau \in P^q(I_m, P_*^p(\mathscr{T}_a))$ such that

$$\begin{aligned}(\sigma_a^\tau, v)_{\mathscr{T}_a,m} - (r_a^\tau, \nabla\cdot v)_{\mathscr{T}_a,m} &= (\psi_a \nabla u_h^\tau, v)_{\mathscr{T}_a,m},\\ &\qquad \forall v \in P^q(I_m, RTN_p^{N,0}(\mathscr{T}_a)),\\ (\nabla\cdot\sigma_a^\tau, q)_{\mathscr{T}_a,m} &= (\psi_a(f - (R_h^\tau)'), q)_{\mathscr{T}_a,m} + (\nabla\psi_a \cdot \nabla u_h^\tau, q)_{\mathscr{T}_a,m},\\ &\qquad \forall q \in P^q(I_m, P_*^p(\mathscr{T}_a)).\end{aligned} \tag{20}$$

Then

$$\sigma_h^\tau = \sum_a \sigma_a^\tau. \tag{21}$$

The reconstructions σ_a^τ and σ_h^τ, exist and satisfy

$$\begin{aligned}0 &= (f - (R_h^\tau)' + \nabla\cdot\sigma_h^\tau, v)_{K,m}\\ &= (f - (u_h^\tau)' + \nabla\cdot\sigma_h^\tau, v)_{K,m} - (\{u_h^\tau\}_{m-1}, v_+^{m-1})_K,\\ &\qquad \forall v \in P^q(I_m, P^p(K)).\end{aligned} \tag{22}$$

4.4 Upper Error Bound

In this section we will present a posteriori upper bound for $\mathrm{EST}(u_h^\tau)$, i.e. we will present the estimate of $\mathrm{EST}(u_h^\tau)$ in terms of data f and u^0, discrete solution u_h^τ (both versions of time discretizations are covered) and functions R_h^τ and σ_h^τ that are derived and easily computable from the discrete solution u_h^τ.

Theorem 4 (Upper error bound) *Let $u \in Y_0$ be the solution of (3) and $u_h^\tau \in X_h^\tau$ be arbitrary. Let R_h^τ be the reconstructions obtained from u_h^τ by (17) and σ_h^τ be the* *reconstruction obtained from u_h^τ by (20) and (21). Then*

$$EST(u_h^\tau) \leq \left(\sum_{K,m} \left(\frac{d_{K,m}}{\pi} \| f - (R_h^\tau)' + \nabla \cdot \sigma_h^\tau \|_{K,m} + \right.\right.$$
$$\left.\left. \frac{d_{K,m}}{h_K} \| \sigma_h^\tau - \nabla u_h^\tau \|_{K,m} + \frac{d_{K,m}}{\tau_m} \| (R_h^\tau - u_h^\tau)' \|_{K,m} \right)^2 \right)^{1/2} \tag{23}$$

The proof of Theorem 4 is a straightforward application of (17) and (22), but it is quite long. For this reason we omit it.

4.5 Asymptotic Lower Error Bound

The goal of this section is to show that the local individual terms from a posteriori estimate (23) are locally effective, i.e. provide a local lower bound to $\mathrm{EST}(u_h^\tau)$, at least in asymptotic sense.

To be able to apply the result in a local way, we need following notation. Let $\mathscr{T}_K$ be a patch consisting of elements surrounding K and K itself. Let $M \subset \overline{\Omega}$, e.g. $M = K$ or $M = \mathscr{T}_K$. We define local version of space Y^τ

$$Y_{M,m}^\tau = \{ v \in Y^\tau : \mathrm{supp}(v) \subset \overline{M \times I_m} \}, \tag{24}$$

and local version of $\mathrm{EST}(w)$

$$\mathrm{EST}_{M,m}(w) = \sup_{0 \neq v \in Y_{M,m}^\tau} \frac{1}{\|v\|_Z} \left(\sum_{K,m} (f, v)_{K,m} - (w', v)_{K,m} \right. \tag{25}$$
$$\left. - (\nabla w, \nabla v)_{K,m} - (\{w\}_{m-1}, v_+^{m-1})_K \right).$$

For the purpose of the effectivity analysis let us assume that f is a space–time polynomial. Otherwise, it is necessary to deal with the classical oscillation term.

Theorem 5 (Local effectivity estimate) *Let $u \in Y_0$ be the solution of (3) and $u_h^\tau \in X_h^\tau$ be arbitrary. Let R_h^τ be the reconstructions obtained from u_h^τ by (17) and σ_h^τ be the reconstruction obtained from u_h^τ by (20) and (21). Let f be a space–time* *polynomial. Then there exists a constant* $C > 0$ *such that*

$$\begin{aligned} &d_{K,m}^2 \|f - (R_h^\tau)' - \nabla \cdot \sigma_h^\tau\|_{K,m}^2 + \frac{d_{K,m}^2}{\tau_m^2}\|R_h^\tau - u_h^\tau\|_{K,m}^2 \\ &+\frac{d_{K,m}^2}{h_K^2}\|\sigma_h^\tau - \nabla u_h^\tau\|_{K,m}^2 \le C\,EST_{\mathscr{T}_K,m}(u_h^\tau)^2. \end{aligned} \tag{26}$$

The proof of Theorem 5 is very technical and quite long. For these reasons we shall skip it in this paper.

Acknowledgements This research was supported by the project GA UK No. 92315 of the Charles University in Prague (F. Roskovec) and by the Grant No. 13-00522S of the Czech Science Foundation (V. Dolejší, M. Vlasák).

References

1. G. Akrivis, C. Makridakis, R.H. Nochetto, Optimal order a posteriori error estimates for a class of Runge-Kutta and Galerkin methods. Numer. Math. **114**(1), 133–160 (2009)
2. G. Akrivis, C. Makridakis, R.H. Nochetto, Galerkin and Runge-Kutta methods: unified formulation, a posteriori error estimates and nodal superconvergence. Numer. Math. **118**(3), 429–456 (2011)
3. D. Braess, J. Schöberl, Equilibrated residual error estimator for edge elements. Math. Comput. **77**(262), 651–672 (2008)
4. V. Dolejší, A. Ern, M. Vohralík., A framework for robust a posteriori error control in unsteady nonlinear advection-diffusion problems. SIAM J. Numer. Anal. **51**(2), 773–793 (2013)
5. A. Ern, M. Vohralík. Polynomial-degree-robust a posteriori estimates in a unified setting for conforming, nonconforming, discontinuous Galerkin, and mixed discretizations. SIAM J. Numer. Anal. **53**(2), 1058–1081 (2015)
6. E. Hairer, S.P. Norsett, G. Wanner. *Solving Ordinary Differential Equations I, Nonstiff Problems*. Springer Series in Computational Mathematics, vol. 8 (Springer, Berlin/Heidelberg/New York, 2000)
7. E. Hairer, G. Wanner. *Solving Ordinary Differential Equations II, Stiff and Differential-Algebraic Problems* (Springer, Berlin, 2002)
8. B.L. Hulme One-step piecewise polynomial Galerkin methods for initial value problems. Math. Comput. **26**, 415–426 (1972)

Part IV
Numerical Linear Algebra and High Performance Computing

Multigrid at Scale?

Mark Ainsworth and Christian Glusa

Abstract The reduced reliability of next generation exascale systems means that the resiliency properties of a numerical algorithm will become an important factor in both the choice of algorithm, and in its analysis. The multigrid algorithm is the workhorse for the distributed solution of linear systems but little is known about its resiliency properties and convergence behavior in a fault-prone environment. In the current work, we propose a probabilistic model for the effect of faults involving random diagonal matrices. We summarize results of the theoretical analysis of the model for the rate of convergence of fault-prone multigrid methods which show that the standard multigrid method will not be resilient. Finally, we present a modification of the standard multigrid algorithm that will be resilient.

1 Introduction

Exascale computing is anticipated to have a huge impact on computational simulation. However, as the number of components in a system becomes larger, the likelihood of one or more components failing or functioning abnormally during an application run increases. The problem is exacerbated by the decreasing physical size of basic components such as transistors, and the accompanying increased possibility of quantum tunneling corrupting logic states [7, 8].

Current day petascale systems already exhibit a diverse range of faults that may occur during computation. These faults can arise from failures in the physical components of the system, or intermittent software faults that appear only in certain application states. One source of faults is cosmic radiation with charged particles, which can lead to memory bit-flips or incorrect behavior of logic units. Future HPC systems are expected to be built from even larger numbers of components than current systems, and the rate of faults in the system will increase accordingly. It is generally accepted that future large-scale systems must operate within a 20 MW power envelope. This will require the usage of lower voltage logic thresholds.

M. Ainsworth (✉) • C. Glusa
Division for Applied Mathematics, Brown University, Providence, RI, USA
e-mail: mark_ainsworth@brown.edu; christian_glusa@brown.edu

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_24

Moreover, cost constraints will result in greater utilization of consumer grade components, with accompanying reduced reliability [8].

Roughly speaking, faults can be classified as follows [3]: *hard* or *stop-fail* faults are faults which would otherwise lead to an immediate program termination, unless treated on the system level. *Soft* faults are those leading to program or data corruption, and which might only result in an erroneous program termination after some delay.

Reported fault rates seem to vary significantly from system to system. On current machines, hard faults have been reported as often as every 4–8 h on the Blue Waters system [8], and (detected) L1-cache soft errors as often as every 5 h on a large BlueGene/L system [9]. The next-generation supercomputers could have a mean-time to failure of about 30 min [24].

Many of the existing algorithms in use today were derived and analyzed without taking account of the effect of these kinds of faults. We believe that the dawning of the exascale era poses new, and exciting, challenges to the numerical analyst in understanding and analyzing the behavior of numerical algorithms on a fault-prone architecture. Our view is that on future exascale systems, the possible impact of faults on the performance of a numerical algorithm must be taken fully into account in the analysis of the method.

In order to alleviate the impact of faults and ensure resilience in a fault-prone environment, several techniques have been proposed and implemented in various parts of the hardware-software stack. Checkpointing on the system and the application level as well as replication of critical program sections are widely used [6, 8, 17]. These techniques can be coupled with statistical analysis, fault models, and hardware health data [8]. On the application level, Algorithm-Based Fault Tolerance (ABFT) describes techniques that duplicate application data to create redundancy [18]. ABFT has been explored in the context of sparse linear algebra [22, 23], and specifically for matrix-vector products in stationary iterative solvers [9, 11, 12, 19, 25]. All methods have in common that a balance needs to be struck between protecting against corruption of results and keeping the overhead reasonable.

The multigrid method is the workhorse for distributed solution of linear systems but little is known about its resiliency properties and convergence behavior in a fault-prone environment. The current article presents a summary of our recent work addressing this problem [1].

The outline of the remainder of this article is as follows: We give a short introduction to multi-level methods in Sect. 2. In Sect. 3, we introduce a model for faults and show simulations of the convergence behavior of a fault-prone two-level method for a finite element method. Section 4 is dedicated to the analysis of products of random matrices and its application to stationary linear iterative methods. Finally, we apply the framework to two- and multi-level methods in Sect. 5, give analytic bounds on the convergence rate, and illustrate their behavior with further simulations. We refer the interested reader for further details and proofs to the articles [1, 2].

2 Multi-level Methods

Let $\Omega \subset \mathbb{R}^d$ be a polygonal domain and set $V := H_0^1(\Omega)$. Starting from an initial triangulation $\mathscr{T}_0$ of Ω into simplices, we obtain $\mathscr{T}_l$ through uniform refinement of $\mathscr{T}_{l-1}$. We define the finite element spaces $V_l := \{v \in H_0^1(\Omega) \cap C\left(\bar{\Omega}\right)$ such that $v\big|_K \in \mathbb{P}_1(K)\,,\ \forall K \in \mathscr{T}_l\}$, and set $n_l := \dim V_l$. For $f \in H^{-1}(\Omega)$, consider the well-posed problem:

$$\text{Find } u \in V \text{ such that} \quad a(u,v) = L(v), \quad \forall v \in V,$$

where $a(u,v) = \int_\Omega \nabla u \cdot \nabla v$ and $L(v) = \int_\Omega f v$. The discretized problem is:

$$\text{Find } u \in V_l \text{ such that} \quad a(u,v) = L(v), \quad \forall v \in V_l.$$

Let $\phi_l^{(i)}$ for $i = 1, \ldots, n_l$ be the global shape function basis of V_l, and ϕ_l the vector of global shape functions. Then the stiffness matrix and the load vector are defined as $A_l := a(\phi_l, \phi_l)$ and $b_l := L(\phi_l)$, so that the problem becomes:

$$\text{Find } u = \phi_l \cdot x_l \in V_l \text{ such that} \quad A_l x_l = b_l. \tag{1}$$

Since $V_{l-1} \subset V_l$, there exists a restriction matrix r_{l+1}^l satisfying $\phi_l = r_{l+1}^l \phi_{l+1}$ along with the corresponding prolongation matrix $p_l^{l+1} = \left(r_{l+1}^l\right)^T$. In particular, this means that the stiffness matrix on level l can be expressed in terms of the matrix at level $l+1$:

$$A_l = a\left(\phi_l, \phi_l\right) = r_{l+1}^l a\left(\phi_{l+1}, \phi_{l+1}\right) p_l^{l+1} = r_{l+1}^l A_{l+1} p_l^{l+1}.$$

We shall omit the sub- and superscripts on r and p whenever it is clear which operator is meant. We shall consider solving the system (1) using the multigrid method[5, 15, 16, 21, 26]. The coarse-grid correction is given by $x_l \leftarrow x_l + pA_{l-1}^{-1} r\left(b_l - A_l x_l\right)$, and has iteration matrix $C_l := I - pA_{l-1}^{-1} r A_l$, while the damped Jacobi smoother is given by $S_l = I - \theta D_l^{-1} A_l$, where D_l is the diagonal of A_l and θ is the relaxation parameter. The multi-level method for the solution of $A_L x_L = b_L$ is given in Algorithm 1. Here, ν_1 and ν_2 are the number of pre- and post-smoothing steps, and γ is the number of coarse-grid corrections.

For the analysis we will use the spectral norm $\|Z\|_2 := \rho\left(ZZ^T\right)^{\frac{1}{2}}$ as well as the energy norm $\|Z\|_A = \left\|A_l^{\frac{1}{2}} Z A_l^{-\frac{1}{2}}\right\|_2$ for matrices $Z \in \mathbb{R}^{n_l \times n_l}$. The convergence of the multi-level method can be proven using the following classical assumptions (see [16]):

(A1) Smoothing property: There exists $\eta : \mathbb{N} \to \mathbb{R}_{\geq 0}$ satisfying $\lim_{\nu \to \infty} \eta(\nu) = 0$ and such that for all levels l

Algorithm 1 Multi-level method $\mathscr{M}_l$

```
Function $\mathscr{M}_l$(right-hand side $b_l$, initial guess $x_l$)
    if $l = 0$ then return $A_0^{-1} x_0$                    (Exact solve on coarsest grid)
    else
        for $i \leftarrow 1$ to $\nu_1$ do
            $x_l \leftarrow x_l + \theta D_l^{-1}(b_l - A_l x_l)$                    (Pre-smoothing)
        $d_{l-1} \leftarrow r(b_l - A_l x_l)$                    (Restriction to coarser grid)
        $e_{l-1}^{(0)} \leftarrow 0$
        for $j \leftarrow 1$ to $\gamma$ do
            $e_{l-1}^{(j)} \leftarrow \mathscr{M}_{l-1}\left(d_{l-1}, e_{l-1}^{(j-1)}\right)$                    (Solve on coarser grid)
        $x_l \leftarrow x_l + p e_{l-1}^{(\gamma)}$                    (Prolongation to finer grid)
        for $i \leftarrow 1$ to $\nu_2$ do
            $x_l \leftarrow x_l + \theta D_l^{-1}(b_l - A_l x_l)$                    (Post-smoothing)
        return $x_l$
```

$$\left\|A_l S_l^{\nu}\right\|_2 \leq \eta(\nu) \left\|A_l\right\|_2, \quad \nu \geq 0.$$

(A2) Approximation property: There exits a constant C_A such that for all levels l

$$\left\|C_l A_l^{-1}\right\|_2 \leq \frac{C_A}{\left\|A_l\right\|_2}.$$

(A3) The smoother is non-expansive, i.e. $\rho(S_l) = \left\|S_l\right\|_A \leq 1$, and there exists a non-increasing function $C_S : \mathbb{N} \to \mathbb{R}_{\geq 0}$ such that for all levels l and $\nu \geq 1$

$$\left\|S_l^{\nu}\right\|_2 \leq C_S(\nu).$$

(A4) There exist positive constants $\underline{C}_p$ and $\overline{C}_p$ such that for all levels l

$$\underline{C}_p^{-1} \left\|x\right\|_2 \leq \left\|px\right\|_2 \leq \overline{C}_p \left\|x\right\|_2 \quad \forall x \in \mathbb{R}^{n_l}.$$

The iteration matrix of the two-level method is given[16] by $E_{TG,L}(\nu_1, \nu_2) = S_L^{\nu_2} C_L S_L^{\nu_1}$. (A1) and (A2) imply that

$$\rho\left(E_{TG,L}(\nu, 0)\right) \leq C_A \eta(\nu).$$

The bound is independent of L and, for large enough ν, smaller than 1. A similar argument can be made for the case involving pre- and post-smoothing using the projection property of the coarse-grid correction C_L. The convergence of the multi-level method for $\gamma \geq 2$ follows using a perturbation argument [16]. Let $E_L(\nu_1, \nu_2, \gamma)$ be the iteration matrix of the multi-level method with finest level L.

Then

$$\rho\left(E_L\left(\nu_1,\nu_2,\gamma\right)\right) \leq \begin{cases} \frac{\gamma}{\gamma-1}\xi, & \gamma \geq 2, \\ \frac{2}{1+\sqrt{1-4C_*\xi}}\xi, & \gamma = 2, \end{cases}$$

where $\xi = \max_{l\leq L}\left\|E_{TG,l}\left(\nu_2,\nu_1\right)\right\|_2$ and C_* depends on ν_1, ν_2 but not on l. The method converges for sufficiently many smoothing steps.

3 Fault Model

The first issue is to decide on how the effect of a fault should be incorporated into the analysis of the algorithm. The simplest and most convenient course of action if a component is subject to corruption, or fails to return a value, is to overwrite the value by zero. We therefore propose to model the effect of a fault on a vector using a random diagonal matrix $\mathscr{X}$, of the form

$$\mathscr{X} = \begin{pmatrix} \chi_1 & & \\ & \ddots & \\ & & \chi_n \end{pmatrix}, \qquad \chi_i = \begin{cases} 1 & \text{with probability } 1-q, \\ 0 & \text{with probability } q. \end{cases} \tag{2}$$

In particular, if a vector $x \in \mathbb{R}^n$ is subject to faults, then the corrupted version of x is given by $\mathscr{X}x$. If all χ_i are independent, we will call the random matrix a matrix of *component-wise* faults. More generally, we shall make the following assumption on the set $\mathscr{S}$ of all the involved faults matrices $\mathscr{X}$:

(A5) There exist constants $v,\ C_e \geq 0$, and for each $\mathscr{X} \in \mathscr{S}$ there exists $e_{\mathscr{X}} \geq 0$ such that for all $\mathscr{X} \in \mathscr{S}$

(a) $\mathscr{X}$ is a random diagonal matrix.
(b) $\|\mathrm{Var}\left[\mathscr{X}\right]\|_2 = \max_{i,j}\left|\mathrm{Cov}\left[\mathscr{X}_{ii},\mathscr{X}_{jj}\right]\right| \leq v$.
(c) $\mathbb{E}\left[\mathscr{X}\right] = e_{\mathscr{X}}I$.
(d) $|e_{\mathscr{X}}-1| \leq C_e v$.

We will think of v as being small. This means that each of the fault matrices $\mathscr{X}$ is close to the identity matrix with high probability. Obviously, the model for component-wise faults introduced above satisfies these assumptions.

In the remainder of this work, we write random matrices in bold letters. If a symbol appears twice, the two occurrences represent the same random matrix and are therefore dependent. If the power of a random matrix appears, we mean the product of identically distributed independent factors.

In summary, we shall model the application of a fault-prone Jacobi smoother as

$$x_l \leftarrow x_l + \mathscr{X}_l^{(\text{pre/post})}\theta D_l^{-1}\left(b_l - A_l x_l\right),$$

which has the same form as a standard Jacobi smoother in which the iteration matrix has been replaced by a random iteration matrix

$$\boldsymbol{S}_l^{(\text{pre/post})} = I - \mathscr{X}_l^{(\text{pre/post})} \theta D_l^{-1} A_l.$$

Here and in what follows, $\mathscr{X}_l^{(\cdot)}$ are generic fault matrices. Suppose that only the calculation of the update can be faulty, and that the previous iterate is preserved. This could be achieved by writing the local components of the current iterate to non-volatile memory or saving it on an adjacent node. The matrices $\mathscr{X}_l^{(\text{pre/post})}$ and D_l^{-1} commute, so that without loss of generality, we can assume that there is just one fault matrix, because any faults in the calculation of the residual can be included in $\mathscr{X}_l^{(\text{pre/post})}$ as well. Moreover, while the application of D_l^{-1} and A_l to a vector is fault-prone, we assume that the entries of D_l^{-1} and A_l itself are not subject to corruption, since permanent changes to them would effectively make it impossible to converge to the correct solution. The matrix entries are generally computed once and for all, and can be stored in non-volatile memory which is protected against corruption. The low writing speed of NVRAM is not an issue since the matrices are written at most once.

The fault-prone two-level method has iteration matrix

$$\boldsymbol{E}_{\boldsymbol{TG},\boldsymbol{l}}(\nu_1, \nu_2) = \left(\boldsymbol{S}_l^{(\text{post})}\right)^{\nu_2} \boldsymbol{C}_l \left(\boldsymbol{S}_l^{(\text{pre})}\right)^{\nu_1},$$

where

$$\boldsymbol{C}_l = I - \mathscr{X}_l^{(p)} p A_{l-1}^{-1} \mathscr{X}_{l-1}^{(r)} r \mathscr{X}_l^{(A)} A_l.$$

Similar arguments as for the smoother can be used to justify the model of faults for the coarse-grid correction. The fault-prone multi-level algorithm is given in Algorithm 2.

In order to illustrate the effect of the faults on the convergence of the algorithm, we apply the two-level version of Algorithm 2 with one step of pre- and post-smoothing using a damped Jacobi smoother with optimal smoothing parameter $\theta = \frac{2}{3}$ to a piecewise linear discretization of the Poisson problem on a square domain.

The domain is partitioned by a uniform triangulation (Fig. 1), and we inject component-wise faults as given in Eq. (2). We plot the evolution of the residual norm over 30 iterations for varying number of degrees of freedom n_L and different probabilities of faults q in Fig. 2 on page 244. We can see that as q increases, the curves start to fan out, with a slope depending on the number of degrees of freedom n_L.

Algorithm 2 Fault-prone multi-level method $\mathscr{M}_l$

```
Function $\mathscr{M}_l$(right-hand side $b_l$, initial guess $x_l$)
    if $l = 0$ then return $A_0^{-1} x_0$                      (Exact solve on coarsest grid)
    else
        for $i \leftarrow 1$ to $\nu_1$ do
            $x_l \leftarrow x_l + \mathscr{X}_l^{(\text{pre},i)} \theta D_l^{-1} (b_l - A_l x_l)$      (Pre-smoothing)
        $d_{l-1} \leftarrow \mathscr{X}_{l-1}^{(r)} r \mathscr{X}_l^{(A)} (b_l - A_l x_l)$     (Restriction to coarser grid)
        $e_{l-1}^{(0)} \leftarrow 0$
        for $j \leftarrow 1$ to $\gamma$ do
            $e_{l-1}^{(j)} \leftarrow \mathscr{M}_{l-1}\left(d_{l-1}, e_{l-1}^{(j-1)}\right)$        (Solve on coarser grid)
        $x_l \leftarrow x_l + \mathscr{X}_l^{(p)} p e_{l-1}^{(\gamma)}$                 (Prolongation to finer grid)
        for $i \leftarrow 1$ to $\nu_2$ do
            $x_l \leftarrow x_l + \mathscr{X}_l^{(\text{post},i)} \theta D_l^{-1} (b_l - A_l x_l)$     (Post-smoothing)
        return $x_l$
```

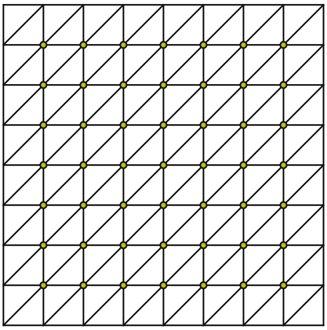


Fig. 1 Mesh for the square domain

4 Lyapunov Exponents

Now, having replaced the iteration matrices by a random quantities, we need to replace the convergence conditions $\rho(E_{TG,l}) < 1$ and $\rho(E_l) < 1$, as the spectral radius has lost its meaning as the asymptotic rate of convergence.

Let $\boldsymbol{A}$ be a random matrix of size $n \times n$ with distribution μ. We will assume that μ is a probability distribution on a finite set $\mathscr{A} \subset GL_n$.

We define the *Lyapunov exponent* $\gamma(\boldsymbol{A})$ as

$$\gamma(\boldsymbol{A}) := \lim_{N\to\infty} \frac{1}{N} \mathbb{E}\left[\log \left\|\boldsymbol{A}^N\right\|\right].$$

(Remember that according to our notation $\boldsymbol{A}^N = \prod_{j=1}^{N} \boldsymbol{A}_j$, where $\boldsymbol{A}_j$ are all independent and of the same distribution as $\boldsymbol{A}$.) This quantity does not depend on the

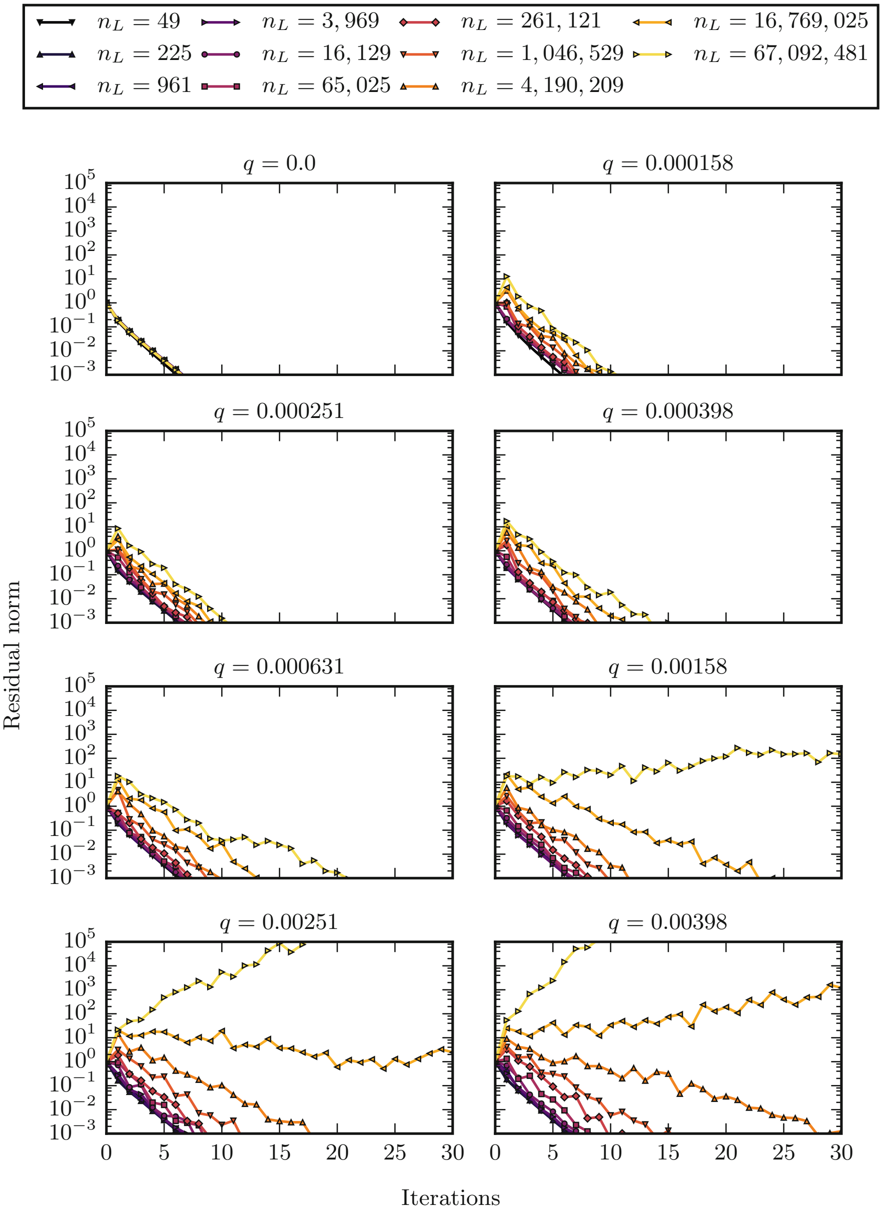


Fig. 2 Evolution of the norm of the residual of the two-level method for the 2d Poisson problem on square domain and component-wise faults in prolongation, restriction, residual and smoother

choice of the norm $\|\cdot\|$ as all norms are equivalent in finite dimension. The definition is motivated by Gelfand's formula $\log \rho(A) = \lim_{N\to\infty} \frac{1}{N} \log \left\|A^N\right\|$ that holds for non-random matrices A.

Furstenberg and Kesten [14] showed that provided that $\mathbb{E}\left[\log (\boldsymbol{A})^{+}\right] < \infty$, $\gamma\left(\boldsymbol{A}\right)$ exists and

$$\gamma\left(\boldsymbol{A}\right) \overset{\text{a.s.}}{=} \lim_{N\to\infty} \frac{1}{N} \log \left\|\boldsymbol{A}^N x_0\right\|.$$

We will instead work with the *Lyapunov spectral radius*

$$\varrho\left(\boldsymbol{A}\right) := e^{\gamma(\boldsymbol{A})} \overset{\text{a.s.}}{=} \lim_{N\to\infty} \left\|\boldsymbol{A}^N x_0\right\|^{\frac{1}{N}}$$

as it is the quantity corresponding to the spectral radius of a non-random matrix. $\varrho\left(\boldsymbol{A}\right) < 1$ means that the product of random matrices is convergent, whereas $\varrho\left(\boldsymbol{A}\right) > 1$ means that it is divergent.

We also define the *generalized Lyapunov exponents* $L(\boldsymbol{A}, \alpha)$ for $\alpha \neq 0$ as

$$L(\boldsymbol{A}, \alpha) := \lim_{N\to\infty} \frac{1}{N} \log \mathbb{E}\left[\left\|\boldsymbol{A}^N\right\|^{\alpha}\right]$$

and the *generalized Lyapunov spectral radii* $\varrho_\alpha\left(\boldsymbol{A}\right)$ as

$$\varrho_\alpha\left(\boldsymbol{A}\right) := e^{\frac{1}{\alpha} L(\boldsymbol{A},\alpha)}.$$

It can be shown (see [4, 10]) that $\frac{1}{N} \log \left\|\boldsymbol{A}^N x_0\right\|$ satisfies a large deviation principle with a rate function I which means that

$$P\left(\frac{1}{N} \log \left\|\boldsymbol{A}^N x_0\right\| = \lambda\right) \approx e^{-NI(\lambda)}.$$

From the above, we find that I satisfies

$$I\left(\gamma\left(\boldsymbol{A}\right)\right) = 0, \qquad I'\left(\gamma\left(\boldsymbol{A}\right)\right) = 0, \qquad I''\left(\gamma\left(\boldsymbol{A}\right)\right) > 0.$$

We can approximately calculate

$$\mathbb{E}\left[\left\|\boldsymbol{A}^N x_0\right\|^{\alpha}\right] = \mathbb{E}\left[e^{N\alpha \frac{1}{N} \log\left\|\boldsymbol{A}^N x_0\right\|}\right] \approx \int e^{N\alpha\lambda} e^{-NI(\lambda)}\, d\lambda \approx e^{N \sup_\lambda \{\alpha\lambda - I(\lambda)\}}$$

and hence $L(\boldsymbol{A}, \alpha) = \sup_\lambda \alpha\lambda - I(\lambda)$. This means that L is the Fenchel-Legendre transform of the rate function I. For each α^*, there is a characteristic growth rate λ^* that corresponds to it, given by $\frac{\partial L}{\partial\alpha}(\alpha^*) = \lambda^*$. Therefore, the values of $L(\boldsymbol{A}, \alpha^*)$ depend on the unlikely sequences $\left\{\boldsymbol{A}^N x_0\right\}_N$ with growth rate λ^* different from $\gamma\left(\boldsymbol{A}\right)$. We have by Jensen's inequality for $\alpha > 0$

$$\varrho_{-\alpha}\left(\boldsymbol{A}\right) \leq \varrho\left(\boldsymbol{A}\right) \leq \varrho_\alpha\left(\boldsymbol{A}\right).$$

Moreover, if A is a non-random matrix, we find $\varrho(A) = \varrho_\alpha(A) = \rho(A)$. The theory of Lyapunov exponents is thoroughly discussed in the book by Bougerol and Lacroix[4].

The following example adapted from [10] shows that the interaction between the values taken by $\boldsymbol{A}$ can be important:

Example 1 Let $a \in \mathbb{R}$ and let the random matrix $\boldsymbol{A}$ take values

$$\begin{pmatrix} 2 & 0 \\ 0 & \frac{1}{4} \end{pmatrix} \quad \text{with probability } (1-p) \qquad \text{and} \qquad \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \text{with probability } p.$$

Obviously, we have $\varrho(\boldsymbol{A}) = 2 > 1$ for $p = 0$ and $\varrho(\boldsymbol{A}) = 1$ for $p = 1$. It can also be shown by decomposition into cycles that

$$\varrho(\boldsymbol{A}) = \lim_{N\to\infty} \left\|\boldsymbol{A}^N\right\|^{\frac{1}{N}} = 2^{\frac{p-1}{2}} < 1 \qquad \text{for } 0 < p < 1.$$

Intuitively, this is the case since expanding and contracting directions are switched by the second matrix and the contracting factor $\frac{1}{4}$ is stronger than the expanding factor 2. This shows that the Lyapunov spectral radius can be discontinuous with respect to the weights and matrices in the support of $\boldsymbol{A}$. In particular, this means that we cannot conclude from $\rho(E_{TG,l}) < 1$ that $\varrho\left(\boldsymbol{E_{TG,l}}\right) < 1$ for small perturbations.

The next example shows that measuring unlikely sequences can lead to serious over-estimation of the convergence rate:

Example 2 Let $\boldsymbol{A}$ be a random 1×1 matrix taking values 2^a with probability $\frac{1}{a}$ and 1 with probability $\frac{a-1}{a}$. Hence the growth rates of all possible sequences are between 1 and 2^a. Then

$$\varrho(\boldsymbol{A}) = 2, \qquad \varrho_\alpha(\boldsymbol{A}) = \left(\frac{1}{a}2^{a\alpha} + \frac{a-1}{a}\right)^{\frac{1}{\alpha}},$$

and for large a we find $\varrho(\boldsymbol{A}) \ll \varrho_\alpha(\boldsymbol{A})$, $\alpha > 0$.

We can approximate the Lyapunov spectral radius $\varrho(\boldsymbol{A})$ by Monte-Carlo simulations of a trajectory $X_N(\omega) = \prod_{j=1}^N \boldsymbol{A}_j(\omega)x_0$ for normalized random initial vector x_0 and calculating $\varrho(\boldsymbol{A}) \approx \|X_N(\omega)\|^{\frac{1}{N}}$. This makes sense because of the almost sure convergence of the latter towards the Lyapunov spectral radius. In order to avoid over- and underflow of the components of the vector $X_N(\omega)$, we renormalize after every step, i.e. we set

$$\tilde{X}_j := \frac{\boldsymbol{A}_j(\omega)\tilde{X}_{j-1}(\omega)}{\left\|\boldsymbol{A}_j(\omega)\tilde{X}_{j-1}(\omega)\right\|}, \qquad \alpha_j := \left\|\boldsymbol{A}_j(\omega)\tilde{X}_{j-1}(\omega)\right\|,$$

and approximate

$$\varrho(\boldsymbol{A}) \approx \exp \frac{1}{N} \sum_{j=1}^{N} \log \alpha_j .$$

The disadvantage of this method is its poor rate of convergence. While the computational estimation of Lyapunov exponents is straightforward, but slow, their analytic calculation is a hard problem, as shown by Tsitsiklis and Blondel [27]. Techniques for small matrices[13] or small support of the distribution[20], as well as small perturbations with respect to the mean value[10] have been developed.

The iteration matrices given in the previous section have dimension n_l and their support has size that is proportional to 2^{n_l}, and the perturbations with respect to the mean value are large in size. We therefore will have to resort to finding an upper bound for the Lyapunov spectral radius. We already saw that for $\alpha > 0$, the generalized Lyapunov spectral radius majorizes the Lyapunov spectral radius. For positive even values of α, its expression can be greatly simplified:

Lemma 3 (Replica trick [10]) *Let* $\boldsymbol{A}$ *be a random square matrix. Then*

$$\varrho_{2k}(\boldsymbol{A}) = \rho\left(\mathbb{E}\left[\boldsymbol{A}^{\otimes 2k}\right]\right)^{\frac{1}{2k}}$$

for $k \in \mathbb{N}$*, where* $Z^{\otimes k} := \underbrace{Z \otimes Z \otimes \cdots \otimes Z}_{k\ times}$*.*

The attraction of the above expression for the generalized Lyapunov exponent stems from the fact that the mean is taken directly of a random matrix, not of a nonlinear function. Hence the method is well suited for linear perturbations. Since the generalized Lyapunov spectral radius increases with α, we use the smallest value allowable in the Replica trick, $\alpha = 2$, and bound

$$\varrho(\boldsymbol{A}) \leq \sqrt{\rho\left(\mathbb{E}\left[\boldsymbol{A}^{\otimes 2}\right]\right)}.$$

As seen in Example 2, this bound is not necessarily sharp, so it is advisable to compare with Monte-Carlo simulations of the Lyapunov spectral radius.

5 Summary of Results on Convergence

With a framework for the analysis of fault-prone stationary iterations in place, we can give the following results whose proofs can be found in [1, 2].

Theorem 4 *Let* $\partial\Omega \in C^2$ *or* Ω *convex and let* A_l *be the stiffness matrices associated the finite element discretization of a second order elliptic PDE on a* *hierarchy of quasi-uniform meshes, and let*

$$\boldsymbol{E_{TG,L}}\left(\nu_1,\nu_2\right)=\left(\boldsymbol{S}_{\boldsymbol{L}}^{(post)}\right)^{\nu_2}\boldsymbol{C_L}\left(\boldsymbol{S}_{\boldsymbol{L}}^{(pre)}\right)^{\nu_1}$$

be the iteration matrix of the two-level method with component-wise faults of rate q in prolongation, restriction, residual and smoother:

$$\begin{aligned}\boldsymbol{C_L}&=I-\mathcal{X}_L^{(p)}pA_{L-1}^{-1}\mathcal{X}_{L-1}^{(r)}r\mathcal{X}_L^{(A)}A_L,\\ \boldsymbol{S}_{\boldsymbol{L}}^{(pre/post)}&=I-\mathcal{X}_L^{(pre/post,i)}D_L^{-1}A_L.\end{aligned}$$

Assume that Assumptions (A1)–(A4) hold. Then

$$\varrho\left(\boldsymbol{E_{TG,L}}\left(\nu_1,\nu_2\right)\right)\leq\left\|E_{TG,L}\left(\nu_1,\nu_2\right)\right\|_A+C\begin{cases}qn_L^{\frac{4-d}{2d}} & d<4,\\ q\left(\log n_L\right)^{\frac{1}{2}} & d=4,\\ q & d>4,\end{cases}$$

where $E_{TG,L}$ is the unperturbed two-level iteration matrix. C is independent of L and q.

In Fig. 3 (top) on page 249, we plot the estimated rate of convergence of the two-level method for the 2d Poisson problem introduced above. We use 1000 iterations to estimate $\varrho\left(\boldsymbol{E_{TG,L}}\left(1,1\right)\right)$ for component-wise faults with varying probability q and varying problem size n_L. Moreover, we plot the behavior predicted by Theorem 4 and the level of $\varrho\left(\boldsymbol{E_{TG,L}}(1,1)\right)=1$. We can see that their slope matches.

The result also holds for the case of an L-shaped domain and for block-wise faults, provided the size of the blocks is fixed, even though the conditions of Theorem 4 are not satisfied.

The above results indicate that two-level methods without protection of some components can not be used in a fault-prone environment. In order to preserve convergence independent of the number of degrees of freedom, we will have to protect one of the fault-prone operations. The cheapest operations are the restriction and the prolongation. The next result shows that the two-level method converges, if the prolongation is protected.

Theorem 5 *Let*

$$\boldsymbol{E_{TG,L}}\left(\nu_1,\nu_2\right)=\left(\boldsymbol{S}_{\boldsymbol{L}}^{(post)}\right)^{\nu_2}\boldsymbol{C_L}\left(\boldsymbol{S}_{\boldsymbol{L}}^{(pre)}\right)^{\nu_1}$$

with smoother and coarse-grid correction given by

$$\begin{aligned}\boldsymbol{S}_{\boldsymbol{L}}^{(pre/post)}&=I-\mathcal{X}_L^{(pre/post,i)}D_L^{-1}A_L,\\ \boldsymbol{C_L}&=I-pA_{L-1}^{-1}\mathcal{X}_{L-1}^{(r)}r\mathcal{X}_L^{(A)}A_L.\end{aligned}$$

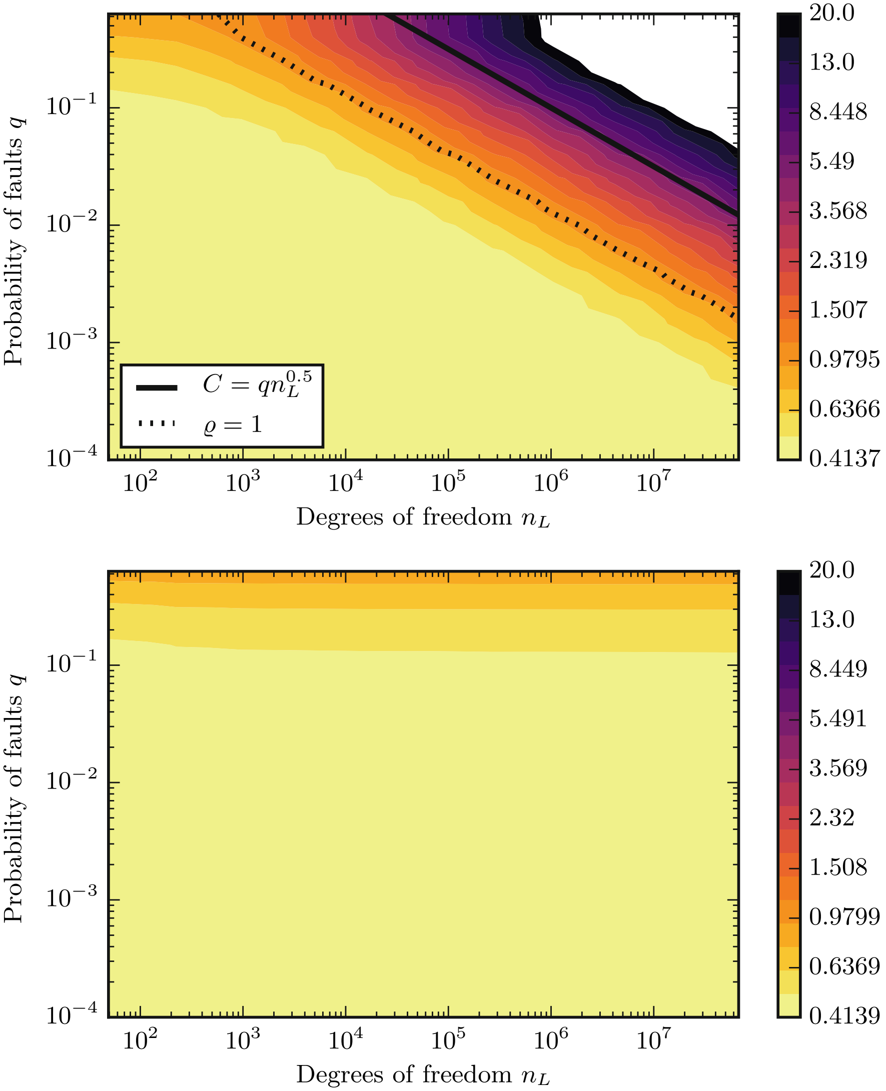


Fig. 3 Asymptotic convergence rate $\varrho\left(\boldsymbol{E}_{TG,L}(1,1)\right)$ of the fault-prone two-level method for the 2d Poisson problem on square domain with component-wise faults in prolongation, restriction, residual and smoother (*top*) and protected prolongation (*bottom*)

Provided Assumptions (A1)–(A5) with

$$\mathscr{S} = \left\{\mathscr{X}_{L-1}^{(r)}, \mathscr{X}_{L}^{(A)}\right\} \cup \left\{\mathscr{X}_{L}^{(pre,i)}\right\}_{i=1}^{\nu_1} \cup \left\{\mathscr{X}_{L}^{(post,i)}\right\}_{i=1}^{\nu_2}$$

hold, we find for any level L that

$$\varrho\left(\boldsymbol{E_{TG,L}}\left(\nu_1, \nu_2\right)\right) \leq \left\| E_{TG,L}\left(\nu_2, \nu_1\right)\right\|_2 + Cv.$$

and C is independent of v and L.

We note that the result holds for more general types of faults including blockwise faults. In Fig. 3 (bottom) on page 249, we plot the rate of convergence of the two-grid method for the already discussed example, this time with protected prolongation. We can see that the rate is essentially independent of the size of the problem, and even is smaller than one for large values of q. The protection can by achieved by standard techniques such as replication. In order to retain performance, the protected prolongation could be overlapped with the application of the post-smoother.

The following theorem shows that the result carries over to the multi-level case:

Theorem 6 *Provided Assumptions (A1)–(A5) with*

$$\mathscr{S} = \bigcup_{l=1}^{L}\left(\left\{\mathscr{X}_{l-1}^{(r)}, \mathscr{X}_{l}^{(A)}\right\} \cup \left\{\mathscr{X}_{l}^{(pre,i)}\right\}_{i=1}^{\nu_1} \cup \left\{\mathscr{X}_{l}^{(post,i)}\right\}_{i=1}^{\nu_2}\right)$$

hold, the number of smoothing steps is sufficient and that v sufficiently small, the perturbed multi-level method converges with a rate bounded by

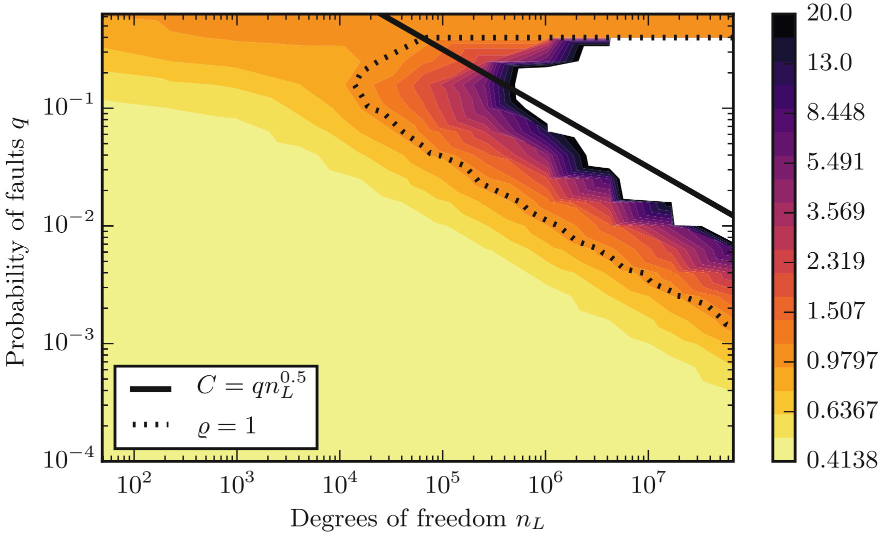
$$\varrho\left(\boldsymbol{E_L}\left(\nu_1, \nu_2, \gamma\right)\right) \leq \begin{cases} \frac{\gamma}{\gamma-1}\xi + Cv, & \gamma \geq 2, \\ \frac{2}{1+\sqrt{1-4C_*\xi}}\xi + Cv, & \gamma = 2, \end{cases}$$

where

$$\xi = \max_{l \leq L} \left\| E_{TG,l}\left(\nu_2, \nu_1\right)\right\|_2,$$

and C_* and C depend on ν_1, ν_2 and the convergence rate of the two-level method, but are independent of L and v.

We also plot the rate of convergence of fault prone multi-level algorithms with two coarse-grid corrections for component-wise faults and protected prolongation in Fig. 4 on page 251, and observe the predicted behavior.



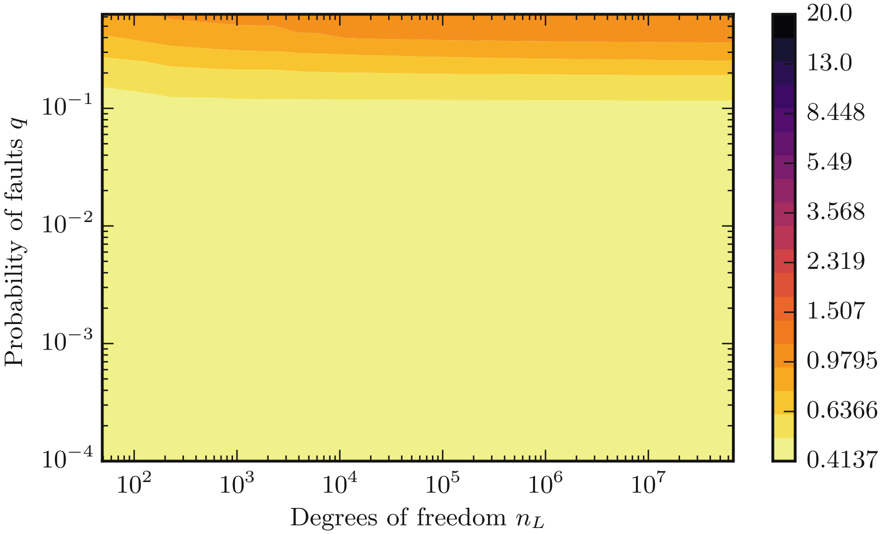


Fig. 4 Asymptotic convergence rate $\varrho\,(\boldsymbol{E_L}(1,1,2))$ of the fault-prone multi-level method for the 2d Poisson problem on square domain with component-wise faults in prolongation, restriction, residual and smoother (*top*) and protected prolongation (*bottom*)

References

1. M. Ainsworth, C. Glusa, Is the multigrid method fault-tolerant? *The Two Grid Case* (Submitted)
2. M. Ainsworth, C. Glusa, Is the multigrid method fault-tolerant? *The Multi Grid Case* (In preparation)
3. A. Avižienis, J.-C. Laprie, B. Randell, C. Landwehr, Basic concepts and taxonomy of dependable and secure computing. IEEE Trans. Dependable Secure Comput. **1**(1), 11–33 (2004)
4. P. Bougerol, J. Lacroix, *Products of Random Matrices with Applications to Schrödinger Operators*. Progress in Probability and Statistics, vol. 8 (Birkhäuser Boston Inc., Boston, 1985)
5. J.H. Bramble, *Multigrid Methods*, vol. 294 (Longman Scientific & Technical, Harlow, 1993)
6. F. Cappello, Fault tolerance in petascale/exascale systems: current knowledge, challenges and research opportunities. Int. J. High Perform. Comput. Appl. **23**(3), 212–226 (2009)
7. F. Cappello, A. Geist, B. Gropp, L. Kale, B. Kramer, M. Snir, Toward exascale resilience. Int. J. High Perform. Comput. Appl. **23**, 374–388 (2009)
8. F. Cappello, A. Geist, W. Gropp, S. Kale, B. Kramer, M. Snir, Toward exascale resilience: 2014 update. Supercomput. Front. Innov. **1**(1), 5–28 (2014)
9. M. Casas, B.R. de Supinski, G. Bronevetsky, M. Schulz, *Fault Resilience of the Algebraic Multi-grid Solver* (ICS'12) (ACM, New York, 2012), pp. 91–100
10. A. Crisanti, G. Paladin, A. Vulpiani, *Products of Random Matrices* (Springer, Berlin/Heidelberg, 1993)
11. T. Cui, J. Xu, C.-S. Zhang, *An Error-Resilient Redundant Subspace Correction Method*, ArXiv e-prints (2013)
12. J. Elliott, F. Mueller, M. Stoyanov, C.G. Webster, *Quantifying the impact of single bit flips on floating point arithmetic*, Technical report ORNL/TM-2013/282, Oak Ridge National Laboratory, 2013
13. M. Embree, L.N. Trefethen, Growth and decay of random Fibonacci sequences. Proc.: Math. Phys. Eng. Sci. **455**(1987), 2471–2485 (1999) (English)
14. H. Furstenberg, H. Kesten, Products of random matrices. Ann. Math. Stat. **31**(2), 457–469 (1960)
15. W. Hackbusch, *Multi-grid Methods and Applications*, vol. 4 (Springer, Berlin, 1985)
16. W. Hackbusch, *Iterative Solution of Large Sparse Systems of Equations*. Applied Mathematical Sciences, vol. 95 (Springer, New York, 1994). Translated and revised from the 1991 German original
17. T. Herault, Y. Robert, *Fault-Tolerance Techniques for High-Performance Computing* (Springer, Cham, 2015)
18. K.-H. Huang, J. Abraham, Algorithm-based fault tolerance for matrix operations. IEEE Trans. Comput. **100**(6), 518–528 (1984)
19. M. Huber, B. Gmeiner, U. Rüde, B. Wohlmuth, *Resilience for multigrid software at the extreme scale*, arXiv preprint arXiv:1506.06185 (2015)
20. R. Mainieri, Zeta function for the Lyapunov exponent of a product of random matrices. Phys. Rev. Lett. **68**, 1965–1968 (1992)
21. S.F. McCormick, W.L. Briggs, V.E. Henson, *A Multigrid Tutorial* (SIAM, Philadelphia, 2000)
22. M. Shantharam, S. Srinivasmurthy, P. Raghavan, *Characterizing the Impact of Soft Errors on Iterative Methods in Scientific Computing* (ICS'11) (ACM, New York, 2011), pp. 152–161
23. J. Sloan, R. Kumar, G. Bronevetsky, Algorithmic approaches to low overhead fault detection for sparse linear algebra, in *2012 42nd Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, Boston (IEEE, 2012), pp. 1–12
24. M. Snir, R.W. Wisniewski, J.A. Abraham, S.V. Adve, S. Bagchi, P. Balaji, J. Belak, P. Bose, F. Cappello, B. Carlson, et al., Addressing failures in exascale computing. Int. J. High Perform. Comput. Appl. **28**(2), 129–173 (2014)

25. M. Stoyanov, C. Webster, Numerical analysis of fixed point algorithms in the presence of hardware faults. SIAM J. Sci. Comput. **37**(5), C532–C553 (2015)
26. U. Trottenberg, C.W. Oosterlee, A. Schüller, *Multigrid* (Academic Press Inc., San Diego, 2001). With contributions by A. Brandt, P. Oswald and K. Stüben
27. J.N. Tsitsiklis, V.D. Blondel, The Lyapunov exponent and joint spectral radius of pairs of matrices are hard-when not impossible-to compute and to approximate. Math. Control Signals Syst. **10**(1), 31–40 (1997) (English)

A Highly Scalable Implementation of Inexact Nonlinear FETI-DP Without Sparse Direct Solvers

Axel Klawonn, Martin Lanser, and Oliver Rheinbach

Abstract A variant of a nonlinear FETI-DP domain decomposition method is considered. It is combined with a parallel algebraic multigrid method (BoomerAMG) in a way which completely removes sparse direct solvers from the algorithm. Scalability to 524,288 MPI ranks is shown for linear elasticity and nonlinear hyperelasticity using more than half of the JUQUEEN supercomputer (JSC, Jülich; TOP500 rank: 11th).

1 Introduction

Classically, nonlinear partial differential equations are solved using a Newton-Krylov approach in which the discretized nonlinear problem is first linearized and then solved by a (possibly globalized) Newton method. In each Newton step, the linear system is then solved iteratively using a Krylov subspace method combined with a scalable preconditioner, e.g., from domain decomposition. In [13, 14], nonlinear FETI-DP (Finite Element Tearing and Interconnecting – Dual Primal) domain decomposition approaches were proposed where the order of the geometrical decomposition and the linearization is interchanged.

Nonlinear domain decomposition as a scalable solution method includes the Additive Schwarz Preconditioned Inexact Newton method [6] (see also [7, 8, 11, 12]) and its recent multiplicative version [20]. Moreover, nonlinear FETI-1 methods [21] and nonlinear Neumann-Neumann methods as a solver [5] are nonlinear domain decomposition methods related to ours.

Versions of nonlinear FETI-DP domain decomposition methods scale to the largest supercomputers currently available, i.e., they have scaled to 524,288 cores

A. Klawonn (✉) • M. Lanser
Mathematisches Institut, Universität zu Köln, Weyertal 86–90, 50931 Köln, Germany
e-mail: axel.klawonn@uni-koeln.de; martin.lanser@uni-koeln.de

O. Rheinbach
Fakultät für Mathematik und Informatik, Institut für Numerische Mathematik und Optimierung,
Technische Universität Bergakademie Freiberg, Akademiestr. 6, 09596 Freiberg, Germany
e-mail: oliver.rheinbach@math.tu-freiberg.de

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_25

[16] and later to the complete Mira supercomputer, i.e., 786,432 cores [15] and 63 billion displacement unknowns in nonlinear hyperelasticity. This is the currently largest range of parallel scalabilty reported for any domain decomposition method. Similarly, but for linear problems, BDDC methods have scaled to 458,752 cores [1]. Algebraic Multigrid (AMG) methods for elasticity have also recently scaled to the same range, i.e., 262,144 cores [2].

In this paper, we consider a new variant of nonlinear FETI-DP domain decomposition methods; see [19]. It is combined with an algebraic multigrid method, completely removing sparse direct solvers from the algorithm.

2 Inexact Nonlinear FETI-DP

The inexact nonlinear FETI-DP approach, first introduced in [19], is a combination of inexact FETI-DP, described in [18], and Nonlinear-FETI-DP-1, introduced in [14]. In this section, we provide a brief description of the method. Our approach is based on the solution of the nonlinear FETI-DP saddle point system

$$\begin{aligned} \widetilde{K}(\tilde{u}) + B^T\lambda - \tilde{f} &= 0 \\ B\tilde{u} \qquad\qquad\quad &= 0 \end{aligned} \tag{1}$$

using Newton's method, which leads to linearized systems of the form

$$\begin{bmatrix} D\widetilde{K}(\tilde{u}) & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} \delta\tilde{u} \\ \delta\lambda \end{bmatrix} = \begin{bmatrix} \widetilde{K}(\tilde{u}) + B^T\lambda - \tilde{f} \\ B\tilde{u} \end{bmatrix}; \tag{2}$$

see also [14] for a detailed description of nonlinear FETI-DP methods.

As in classical linear FETI-DP methods, we assume a nonoverlapping domain decomposition of the computational domain. The resulting interface variables are split into primal (Π) and dual variables (Δ). Variables in the interior part of the subdomains are denoted by I, and we define the index set $B := [I\ \ \Delta]$. The block matrix $D\widetilde{K}(\tilde{u})$ is partially assembled in the primal variables Π, while the remaining part of $D\widetilde{K}(\tilde{u})$ typically is block diagonal. Each diagonal block is associated to one of the FETI-DP subdomains. The operator B is a classical jump operator, well known from any FETI-DP literature. Following the notation in [18], we define

$$\mathscr{A} := \begin{bmatrix} D\widetilde{K}(\tilde{u}) & B^T \\ B & 0 \end{bmatrix}, \quad \mathscr{F} := \begin{bmatrix} \widetilde{K}(\tilde{u}) + B^T\lambda - \tilde{f} \\ B\tilde{u} \end{bmatrix}, \quad \text{and} \quad x := \begin{bmatrix} \delta\tilde{u} \\ \delta\lambda \end{bmatrix}.$$

We apply a Krylov method, e.g., GMRES, to the preconditioned system

$$\mathscr{B}_L^{-1}\mathscr{A}x = \mathscr{B}_L^{-1}\mathscr{F} \tag{3}$$

in order to solve the linearized system in each Newton step. The block triangular preconditioner $\mathscr{B}_L$ is defined by

$$\mathscr{B}_L := \begin{bmatrix} \widehat{K}^{-1} & 0 \\ M^{-1}B\widehat{K}^{-1} & M^{-1} \end{bmatrix},$$

where $\widehat{K}^{-1}$ is a sufficiently good preconditioner for $D\widetilde{K}(\tilde{u})$, and M^{-1} is one of the standard FETI-DP preconditioners. Throughout this paper, the application of $\widehat{K}^{-1}$ consists of one V-cycle of a parallel AMG method applied to the complete system $D\widetilde{K}(\tilde{u})$. We investigate two different choices for the preconditioner M^{-1}. First, we use the standard Dirichlet preconditioner

$$M^{-1} := M^{-1}_{\mathrm{FETI_D}} := \sum_{i=1}^{N} B^{(i)}_{\Delta,D}\, S^{(i)}_{\Delta\Delta}\, B^{(i)T}_{\Delta,D},$$

which is a weighted sum of Schur complements

$$S^{(i)}_{\Delta\Delta} := DK^{(i)}_{\Delta\Delta} - DK^{(i)}_{\Delta I}\ (DK^{(i)}_{II})^{-1}\ DK^{(i)T}_{\Delta I}$$

on the dual part of the interface. Here, the matrices $DK^{(i)}_{II}$, $DK^{(i)}_{\Delta\Delta}$, and $DK^{(i)}_{\Delta I}$ correspond to blocks of the tangential matrix $D\widetilde{K}(\tilde{u})$ and are local to the i-th subdomain.

Second, to completely remove sparse direct solvers from the method, we can replace an application of $\left(DK^{(i)}_{II}\right)^{-1}$ by some cycles of a local AMG method. We denote this modified Dirichlet preconditioner by $M^{-1}_{\mathrm{FETI_{D/AMG}}}$. Let us note that this approach does not guarantee spectral equivalence to the (exact) Dirichlet preconditioner unless the interior system is solved accurately enough. Nevertheless, this modified preconditioner often leads to appropriate results; see also [17]. Finally, the complete algorithm is presented in Fig. 1.

```
**Init:** $\tilde{u}^{(0)} \in \widetilde{W}$
**for** $k = 0, ..., convergence$
    **build:** $\widetilde{K}(\tilde{u}^{(k)})$, $D\widetilde{K}(\tilde{u}^{(k)})$, and $M^{-1}$
    **iterative Krylov solve for** $x = [\delta\tilde{u}^{(k)T}, \delta\lambda^{(k)}]$ **using left**
    **preconditioner** $\mathscr{B}_L^{-1} := \mathscr{B}_L^{-1}(\widehat{K}^{-1}, M^{-1})$**:**
        $\mathscr{A}x = \mathscr{F}$        // (see eq. (3))
    **update:**
    $\tilde{u}^{(k+1)} := \tilde{u}^{(k)} - \delta\tilde{u}^{(k)}$
    $\lambda^{(k+1)} := \lambda^{(k)} - \delta\lambda^{(k)}$
**end**
```

Fig. 1 Pseudocode of the **Inexact-Nonlinear-FETI-DP**. The application of $\widehat{K}^{-1}$ consists of cycles of a parallel AMG method

We have implemented the Inexact-Nonlinear-FETI-DP method in PETSc 3.6.2 [4] using C/C++ and MPI. We decided to implement the matrix $D\widetilde{K}(\tilde{u})$ and the jump operator B as MPI parallel sparse matrices of the type *MPIAIJ*, which is provided by PETSc. All rows of $D\widetilde{K}(\tilde{u})$ corresponding to the interior and interface nodes of the i-th subdomain are distributed to the same MPI rank, i.e., the local subdomain block $\left[DK_{BB}^{(i)}(\tilde{u}) \; D\widetilde{K}_{B\Pi}^{(i)}(\tilde{u}) \right]$ is assigned to one MPI rank. The rows corresponding to the globally assembled FETI-DP coarse space are distributed equally to all MPI ranks, and thus we do not obtain the typical block structure

$$D\widetilde{K}(\tilde{u}) := \begin{bmatrix} DK_{BB}(\tilde{u}) & D\widetilde{K}_{\Pi B}^{T}(\tilde{u}) \\ D\widetilde{K}_{\Pi B}(\tilde{u}) & D\widetilde{K}_{\Pi\Pi}(\tilde{u}) \end{bmatrix}$$

in our implementation. We always try to distribute a primal variable to one of the MPI ranks handling a neighboring subdomain. This strategy should reduce communication. The rows of B^T are distributed equivalently.
As preconditioner for $D\widetilde{K}(\tilde{u})$, we always use one V-cycle of BoomerAMG [9]. Other spectrally equivalent preconditioners are also possible, e.g., multilevel preconditioners from domain decomposition. Although not being spectrally equivalent, preconditioners such als ILU (incomplete LU) or SPAI (Sparse Approximate Inverse) could also be used as long as the local matrices are not too ill-conditioned. In some of our numerical tests, we also use the global matrix (GM) approach introduced in [3] and used in [2], which guarantees the exact interpolation of chosen smooth error vectors, e.g., rigid body modes (rotations and translations). This can improve the quality of AMG as preconditioner for elasticity problems. If using the GM interpolation, we have to provide the rotations on the finite element space $\widetilde{W}$, i.e., the rotation of the coarse space and the subdomain nodes. We also present an algorithmic description of Inexact-Nonlinear-FETI-DP in form of a pseudocode in Fig. 1.

3 Model Problems and Numerical Results

We consider three different elasticity problems. First, we investigate a compressible linear elasticity problem

$$-2\mu \operatorname{div}(\varepsilon(u)) - \lambda \operatorname{grad}(\operatorname{div}(u)) = f$$

with $\lambda = \frac{\nu E}{(1+\nu)(1-2\nu)}, \mu = \frac{E}{2(1+\nu)}$. We choose $E = 210$ and $\nu = 0.3$, consider a rectangular domain $\Omega := [0\,,\,8] \times [0\,,\,1]$ and a homogeneous Dirichlet boundary condition in all nodes $(x, y) \in \Omega$ with $x = 0$. A constant volume force in y-direction is applied to the complete **linear 2D beam**; see Table 1 for some weak scalability results. As a second model problem, we consider the same domain, boundary condition, material parameters, and volume force, but choose a nonlinear

Table 1 **Linear 2D beam**; one V-cycle of BoomerAMG with nodal HMIS coarsening and GM interpolation is used in all cases; **It.** denotes the number of GMRES iterations; the baseline of the parallel efficiency **Eff.** is the fastest time to solution on 32 MPI ranks (1 node)

# MPI ranks	D.o.f.	M^{-1}	Newton It.	It.	Time to solution (s)	Eff. (%)	Time assembly Eq. (2) (s)	Time setup $\hat{K}^{-1}$ (s)	Time setup M^{-1} (s)	Time GMRES (s)
32	1,644,162	$M^{-1}_{\mathrm{FETI_D}}$	1	23	46.3	100.0	5.7	8.6	10.2	20.1
		$M^{-1}_{\mathrm{FETI_{D/AMG}}}$	1	27	51.4	90.1	5.7	8.7	2.9	32.6
128	6,565,122	$M^{-1}_{\mathrm{FETI_D}}$	1	20	45.2	102.4	5.8	8.8	11.4	17.7
		$M^{-1}_{\mathrm{FETI_{D/AMG}}}$	1	25	49.5	93.5	5.8	8.8	2.9	30.4
512	26,237,442	$M^{-1}_{\mathrm{FETI_D}}$	1	18	45.3	102.2	5.9	8.8	11.5	17.5
		$M^{-1}_{\mathrm{FETI_{D/AMG}}}$	1	23	49.5	93.5	5.9	8.9	2.9	30.3
2,048	104,903,682	$M^{-1}_{\mathrm{FETI_D}}$	1	15	41.4	111.8	5.8	8.9	11.6	13.5
		$M^{-1}_{\mathrm{FETI_{D/AMG}}}$	1	22	46.0	100.6	5.8	8.9	3.0	26.9
8,192	419,522,562	$M^{-1}_{\mathrm{FETI_D}}$	1	14	40.9	113.2	5.9	9.1	11.4	12.7
		$M^{-1}_{\mathrm{FETI_{D/AMG}}}$	1	20	44.1	105.0	5.9	9.1	3.0	24.5
32,768	1,677,905,922	$M^{-1}_{\mathrm{FETI_D}}$	1	12	39.9	116.0	6.2	9.3	11.7	11.0
		$M^{-1}_{\mathrm{FETI_{D/AMG}}}$	1	20	44.9	103.1	6.2	9.3	3.0	24.6
131,072	6,711,255,042	$M^{-1}_{\mathrm{FETI_D}}$	1	13	42.2	109.7	6.7	9.9	11.4	11.9
		$M^{-1}_{\mathrm{FETI_{D/AMG}}}$	1	20	46.5	99.6	6.7	9.8	3.0	24.6
524,288	26,844,282,882	$M^{-1}_{\mathrm{FETI_D}}$	1	14	50.5	91.7	9.2	11.1	11.4	13.2
		$M^{-1}_{\mathrm{FETI_{D/AMG}}}$	1	22	56.9	81.4	9.3	11.1	3.3	27.8

Neo-Hooke material. The strain energy density function of the Neo-Hooke material W [10, 22] is given by

$$W(u) = (\mu/2)\big(\mathrm{tr}(\mathbf{F}^T\mathbf{F}) - 3\big) - \mu \ln(J) + (\lambda/2)\ln^2(J)$$

with the deformation gradient $\mathbf{F}(x) := \nabla\boldsymbol{\varphi}(\mathbf{x})$. Here, $\boldsymbol{\varphi}(\mathbf{x}) = \mathbf{x} + \mathbf{u}(\mathbf{x})$ denotes the deformation and $\mathbf{u}(\mathbf{x})$ the displacement of $\mathbf{x}$. We present weak scalability tests for the **nonlinear 2D beam** in Table 2. Our third model problem is strongly heterogeneous. We consider a rectangular domain $[0\,,\,2] \times [0\,,\,1]$ and apply the deformation $\mathbf{F} = \begin{bmatrix} 1.1 & 0 \\ 0 & 1 \end{bmatrix}$ in each boundary node. Again, a Neo-Hooke material with $E = 210$ and $\nu = 0.3$ is used, but we consider one slightly off-centered circular inclusion of stiff material ($E = 210{,}000$ and $\nu = 0.3$) in each FETI-DP subdomain. The weak scalability results for the **heterogeneous nonlinear problem** are presented in Table 3. Let us remark that we always use a single square FETI-DP subdomain per MPI rank and consider all vertices to be primal. We choose a discretization with piecewise quadratic triangular finite elements in all our experiments. All computations are performed on JUQUEEN BlueGene/Q at Forschungszentrum Jülich using 32 MPI ranks per node. JUQUEEN is currently ranked 11th in the TOP500 list of world's fastest supercomputers. We always choose HMIS coarsening and ext+i interpolations from BoomerAMG. The GM approach is used additionally in the 2D beam computations and hybrid AMG (nodal coarsening and unknown based interpolation) for the heterogeneous Neo-Hooke problem. The choice of the AMG components is motivated by our experience gained in [2]. We use UMFPACK for $M^{-1}_{\mathrm{FETI_D}}$.

The total runtime of Inexact-Nonlinear-FETI-DP basically splits into four different phases: the assembly of all parts of the saddle point system in (2) (including the assembly of the rigid body modes for the GM approach), the setup of the Dirichlet preconditioner M^{-1}, the BoomerAMG setup time to create $\hat{K}^{-1}$, and finally the iterative solution using preconditioned GMRES. Runtimes for these four phases are presented in all tables. Since the setup of M^{-1} always scales nearly perfectly, we will only discuss the remaining timings. Let us remark that the setup of $M^{-1}_{\mathrm{FETI_{D/AMG}}}$ is up to four times faster than the setup of $M^{-1}_{\mathrm{FETI_D}}$, since direct factorizations are avoided. In contrast, since an application of an AMG V-cycle is more expensive than a forward-backward solve in UMFPACK, one preconditioned GMRES iteration using $M^{-1}_{\mathrm{FETI_D}}$ is cheaper.

For the linear elastic beam (Table 1), we obtain weak scalability with more than 90 % parallel efficiency. Efficiencies higher than 100 % result from numerical effects, i.e., a decreasing number of GMRES iterations. These compensate some inefficiencies in the assembly of the saddle point system and the BoomerAMG setup. The direct solver-free $M^{-1}_{\mathrm{FETI_{D/AMG}}}$ also convinces in runtime, and numerical as well as parallel scalability for this model problem. Similar observations can be made for the nonlinear beam (Table 2). In this case, scalability is less optimal caused by a more expensive AMG setup. However, obtaining a parallel efficiency of more than 56 % scaling from 32 to 524,288 MPI ranks is still remarkable, especially, since the total problem sizes are smaller. For the heterogeneous material (Table 3),

Table 2 **Nonlinear 2D beam**; one V-cycle of BoomerAMG with nodal HMIS coarsening and GM interpolation is used in all cases; **It.** denotes the number of GMRES iterations; the baseline of the parallel efficiency **Eff.** is the fastest time to solution on 32 MPI ranks (1 node); all values are summed values over the 4 Newton steps

					Time		Time	Time	Time	
# MPI			Newton		to		assembly	setup	setup	Time
ranks	D.o.f.	M^{-1}	It.	It.	solution (s)	Eff. (%)	Eq. (2) (s)	$\hat{K}^{-1}$ (s)	M^{-1} (s)	GMRES (s)
32	643,602	$M^{-1}_{\mathrm{FETI_D}}$	4	66	53.4	100.0	5.7	14.3	11.7	21.0
		$M^{-1}_{\mathrm{FETI_{D/AMG}}}$	4	70	56.7	94.2	5.7	14.2	4.4	31.5
128	2,567,202	$M^{-1}_{\mathrm{FETI_D}}$	4	65	55.1	96.9	5.8	14.8	12.6	21.0
		$M^{-1}_{\mathrm{FETI_{D/AMG}}}$	4	72	58.8	90.8	6.0	14.8	4.4	32.7
512	10,254,402	$M^{-1}_{\mathrm{FETI_D}}$	4	57	55.2	96.7	6.0	17.0	12.6	18.7
		$M^{-1}_{\mathrm{FETI_{D/AMG}}}$	4	66	58.5	91.3	6.1	17.0	4.4	30.2
2,048	40,988,802	$M^{-1}_{\mathrm{FETI_D}}$	4	52	54.5	98.0	6.1	17.6	12.7	17.2
		$M^{-1}_{\mathrm{FETI_{D/AMG}}}$	4	64	58.6	91.1	6.1	17.6	4.4	29.4
8,192	163,897,602	$M^{-1}_{\mathrm{FETI_D}}$	4	52	55.9	95.5	6.4	18.2	12.7	17.7
		$M^{-1}_{\mathrm{FETI_{D/AMG}}}$	4	64	59.5	89.8	6.4	18.2	4.5	29.5
32,768	655,475,202	$M^{-1}_{\mathrm{FETI_D}}$	4	52	57.6	92.7	6.7	19.2	12.7	17.7
		$M^{-1}_{\mathrm{FETI_{D/AMG}}}$	4	70	64.6	82.7	6.7	19.2	4.5	32.8
131,072	2,621,670,402	$M^{-1}_{\mathrm{FETI_D}}$	4	44	60.4	88.4	8.6	21.8	12.8	15.0
		$M^{-1}_{\mathrm{FETI_{D/AMG}}}$	4	61	65.7	81.3	8.5	21.6	4.7	28.5
524,288	10,486,220,802	$M^{-1}_{\mathrm{FETI_D}}$	4	47	86.6	61.7	17.6	28.6	14.0	17.1
		$M^{-1}_{\mathrm{FETI_{D/AMG}}}$	4	66	94.0	56.8	17.8	28.6	5.7	32.5

Table 3 **Heterogeneous Neo-Hooke problem**; one V-cycle of BoomerAMG with nodal HMIS coarsening is used in all cases; **It.** denotes the number of GMRES iterations; the baseline of the parallel efficiency **Eff.** is the fastest time to solution on 32 MPI ranks (1 node); all values are summed values over the 4 Newton steps

# MPI ranks	D.o.f.	M^{-1}	Newton It.	It.	Time to solution (s)	Eff. (%)	Time assembly Eq. (2) (s)	Time setup $\hat{K}^{-1}$ (s)	Time setup M^{-1} (s)	Time GMRES (s)
32	642,402	$M^{-1}_{\mathrm{FETI\text{-}D}}$	4	90	45.6	100.0	5.7	3.6	11.7	23.7
		$M^{-1}_{\mathrm{FETI\text{-}D/AMG}}$	4	98	52.2	87.4	5.8	3.6	3.7	38.2
128	2,564,802	$M^{-1}_{\mathrm{FETI\text{-}D}}$	4	110	52.3	87.2	6.0	3.6	12.3	29.4
		$M^{-1}_{\mathrm{FETI\text{-}D/AMG}}$	4	105	55.6	82.0	6.0	3.6	3.7	41.3
512	10,249,602	$M^{-1}_{\mathrm{FETI\text{-}D}}$	4	120	55.8	81.7	6.0	3.8	12.6	32.6
		$M^{-1}_{\mathrm{FETI\text{-}D/AMG}}$	4	111	58.6	77.8	6.0	3.8	3.7	44.1
2,048	40,979,202	$M^{-1}_{\mathrm{FETI\text{-}D}}$	4	129	58.7	77.7	6.0	3.9	12.6	35.1
		$M^{-1}_{\mathrm{FETI\text{-}D/AMG}}$	4	117	61.6	74.0	6.2	4.0	3.8	46.7
8,192	163,878,402	$M^{-1}_{\mathrm{FETI\text{-}D}}$	4	147	64.2	71.0	6.2	4.2	12.7	40.2
		$M^{-1}_{\mathrm{FETI\text{-}D/AMG}}$	4	132	68.1	67.0	6.4	4.2	3.8	52.6
32,768	655,436,802	$M^{-1}_{\mathrm{FETI\text{-}D}}$	4	156	68.1	67.0	6.5	4.6	12.7	43.0
		$M^{-1}_{\mathrm{FETI\text{-}D/AMG}}$	4	135	70.5	64.7	6.9	4.6	3.8	54.1
131,072	2,621,593,602	$M^{-1}_{\mathrm{FETI\text{-}D}}$	4	180	79.7	57.2	8.5	5.6	12.8	50.2
		$M^{-1}_{\mathrm{FETI\text{-}D/AMG}}$	4	159	85.0	53.7	8.6	5.5	4.0	64.6
524,288	10,486,220,802	$M^{-1}_{\mathrm{FETI\text{-}D}}$	4	189	104.7	43.6	17.5	7.8	14.0	56.1
		$M^{-1}_{\mathrm{FETI\text{-}D/AMG}}$	4	177	114.7	39.8	17.4	8.0	5.0	75.0

the parallel scalability suffers from a certain loss in numerical scalability. Note that the number of heterogeneities increases with the number of ranks. Let us summarize that Inexact-Nonlinear-FETI-DP METHODS are robust for different homogeneous and heterogeneous elasticity problems. Also the variant without sparse direct solvers performS well. All components of the method show sufficient scalability.

Acknowledgements This work was supported in part by the German Research Foundation (DFG) through the Priority Programme 1648 "Software for Exascale Computing" (**SPPEXA**) under KL 2094/4 and RH 122/2. The authors gratefully acknowledge the Gauss Centre for Supercomputing (GCS) for providing computing time through the John von Neumann Institute for Computing (NIC) on the GCS share of the supercomputer JUQUEEN at Jülich Supercomputing Centre.

References

1. S. Badia, A.F. Martin, J. Principe, Multilevel balancing domain decomposition at extreme scales. SIAM J. Sci. Comput. 38(1), C22ŰC52 (2016)
2. A.H. Baker, A. Klawonn, T. Kolev, M. Lanser, O. Rheinbach, U.M. Yang, *Scalability of Classical Algebraic Multigrid for Elasticity to Half a Million Parallel Tasks*. Lecture Notes in Computational Science and Engineering (2015). TUBAF Preprint: 2015–14. http://tu-freiberg.de/fakult1/forschung/preprints
3. A.H. Baker, T.V. Kolev, U.M. Yang, Improving algebraic multigrid interpolation operators for linear elasticity problems. Numer. Linear Algebra Appl. **17**(2–3), 495–517 (2010)
4. S. Balay, W.D. Gropp, L.C. McInnes, B.F. Smith, Efficient management of parallelism in object oriented numerical software libraries, in *Modern Software Tools in Scientific Computing*, ed. by E. Arge, A.M. Bruaset, H.P. Langtangen (Birkhauser Press, Boston, 1997), pp. 163–202
5. F. Bordeu, P.-A. Boucard, P. Gosselet, Balancing domain decomposition with nonlinear relocalization: parallel implementation for laminates, in *Proceedings of the 1st International Conference on Parallel, Distributed and Grid Computing for Engineering*, ed. by P.I. B.H.V. Topping (Civil-Comp Press, Stirlingshire, 2009)
6. X.-C. Cai, D.E. Keyes, Nonlinearly preconditioned inexact Newton algorithms. SIAM J. Sci. Comput. **24**(1), 183–200 (2002) (electronic)
7. X.-C. Cai, D.E. Keyes, L. Marcinkowski, Non-linear additive Schwarz preconditioners and application in computational fluid dynamics. Int. J. Numer. Methods Fluids **40**(12), 1463–1470 (2002)
8. C. Groß, R. Krause, On the globalization of aspin employing trust-region control strategies – convergence analysis and numerical examples. Technical report 2011–03, Institute of Computational Science, Universita della Svizzera italiana (2011)
9. V.E. Henson, U.M. Yang, BoomerAMG: a parallel algebraic multigrid solver and preconditioner. Appl. Numer. Math. **41**, 155–177 (2002)
10. G.A. Holzapfel, *Nonlinear Solid Mechanics. A Continuum Approach for Engineering* (Wiley, Chichester, 2000)
11. F.-N. Hwang, X.-C. Cai, Improving robustness and parallel scalability of Newton method through nonlinear preconditioning, in *Domain Decomposition Methods in Science and Engineering*, ed. by R. Kornhuber, R.W. Hoppe, J. Périaux, O. Pironneau, O. Widlund, J. Xu. Lecture Notes in Computational Science and Engineering, vol. 40 (Springer, Berlin, 2005), pp. 201–208
12. F.-N. Hwang, X.-C. Cai, A class of parallel two-level nonlinear Schwarz preconditioned inexact Newton algorithms. Comput. Methods Appl. Mech. Eng. **196**(8), 1603–1611 (2007)

13. A. Klawonn, M. Lanser, P. Radtke, O. Rheinbach, in *On an Adaptive Coarse Space and on Nonlinear Domain Decomposition.* , ed. by J. Erhel, M.J. Gander, L. Halpern, G. Pichot, T. Sassi, O.B. Widlund. Lecture Notes in Computational Science and Engineering, vol. 98 (Springer International Publishing, Switzerland, 2014), pp. 71–83
14. A. Klawonn, M. Lanser, O. Rheinbach, Nonlinear FETI-DP and BDDC methods. SIAM J. Sci. Comput. **36**(2), A737–A765 (2014)
15. A. Klawonn, M. Lanser, O. Rheinbach, FE2TI: computational scale bridging for dual-phase steels, in *Parallel Computing: On the Road to Exascale*. Advances in Parallel Computing, vol. 27 (IOS Press, Amsterdam, 2015), pp. 797–806
16. A. Klawonn, M. Lanser, O. Rheinbach, Toward extremely scalable nonlinear domain decomposition methods for elliptic partial differential equations. SIAM J. Sci. Comput. **37**(6), C667–C696 (2015)
17. A. Klawonn, L.F. Pavarino, O. Rheinbach, Spectral element FETI-DP and BDDC preconditioners with multi-element subdomains. Comput. Meth. Appl. Mech. Eng. **198**, 511–523 (2008)
18. A. Klawonn, O. Rheinbach, Inexact FETI-DP methods. Int. J. Numer. Methods Eng. **69**(2), 284–307 (2007)
19. M. Lanser, Nonlinear FETI-DP and BDDC methods. Ph.D. thesis, Universität zu Köln (2015)
20. L. Liu, D.E. Keyes, Field-split preconditioned inexact Newton algorithms. SIAM J. Sci. Comput. **37**(3), A1388–A1409 (2015)
21. J. Pebrel, C. Rey, P. Gosselet, A nonlinear dual-domain decomposition method: application to structural problems with damage. Int. J. Multiscale Comput. Eng. **6**(3), 251–262 (2008)
22. O. Zienkiewicz, R. Taylor, *The Finite Element Method for Solid and Structural Mechanics* (Elsevier, Oxford, 2005)

A Parallel Multigrid Solver for Time-Periodic Incompressible Navier–Stokes Equations in 3D

Pietro Benedusi, Daniel Hupp, Peter Arbenz, and Rolf Krause

Abstract We present a parallel and efficient multilevel solution method for the nonlinear systems arising from the discretization of Navier–Stokes (N-S) equations with finite differences. In particular we study the incompressible, unsteady N-S equations with periodic boundary condition in time. A sequential time integration limits the parallelism of the solver to the spatial variables and can therefore be an obstacle to parallel scalability. Time periodicity allows for a space-time discretization, which adds time as an additional direction for parallelism and thus can improve parallel scalability. To achieve fast convergence, we used a space-time multigrid algorithm with a SCGS smoothing procedure (symmetrical coupled Gauss–Seidel, a.k.a. box smoothing). This technique, proposed by Vanka (J Comput Phys 65:138–156, 1986), for the steady viscous incompressible Navier–Stokes equations is extended to the unsteady case and its properties are studied using local Fourier analysis. We used numerical experiments to analyze the scalability and the convergence of the solver, focusing on the case of a pulsatile flow.

1 Introduction

The study of time periodic problems in computational fluid dynamics (CFD) is interesting for multiple reasons.

Pulsatile flows, for example, are present in a variety of physical problems as the modeling of biological fluids in living creatures (biofluid mechanics). Examples in this field are the blood flowing in veins [3, 4] or air in human lungs [5].

P. Benedusi (✉) • R. Krause
Institute of Computational Science, USI, Lugano, Switzerland
e-mail: pietro.benedusi@usi.ch; rolf.krause@usi.ch

D. Hupp • P. Arbenz
Computer Science Department, ETHZ, Zürich, Switzerland
e-mail: huppd@inf.ethz.ch; arbenz@inf.ethz.ch

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_26

Moreover it is well known that today high performance computing systems are typically massively parallel, i.e. they consist of a large number of computer nodes (providing CPUs or GPUs) connected by a high speed network. Thus, algorithms with high concurrency must be developed to efficiently exploit these resources. Traditionally, sequential time integration schemes are a bottleneck for the scalability of the solving process. In fact, when time stepping methods are used for evolutionary problems the corresponding algorithm is inherently sequential. If parallelization takes place only in space (e.g. trough domain decomposition), the computation time can only be reduced to a certain limit. When the number of degrees of freedom per core is "small" the spatial decomposition is saturated and communication takes over.

Thus it is convenient to employ also parallelism in time. The time periodicity modifies the problem structure and the design of time-parallel solution strategies becomes natural if a full space-time parallel algorithm is used. To get an optimal convergence we construct a space-time multigrid solver with a SCGS smoother [2]. The nonlinearity of the problem will be treated a priori, using a Picard iterative process. For the discretization of the Navier–Stokes equations, in $3 + 1$ dimensions, we use fourth and second order finite differences on a staggered grid. The final goal will be to investigate the convergence properties and the strong/weak scaling of the parallel solver.

2 Governing Equations

In three spatial dimensions, the unsteady incompressible Navier–Stokes (N-S) equations in the primitive variables $(\mathbf{u}, p)$ can be written as follows, with $\mathbf{u} = \mathbf{u}(\mathbf{x}, t)$ the velocity vector field, $p = p(\mathbf{x}, t)$ the pressure scalar field, $\Omega \subseteq \mathbb{R}^3$ a domain with a Lipschitz boundary and $\mathrm{Re}, \alpha \in \mathbb{R}^+$:

$$\alpha^2 \partial_t \mathbf{u} + \mathrm{Re}(\mathbf{u} \cdot \nabla)\mathbf{u} = -\mathrm{Re}\nabla p + \triangle \mathbf{u} + \mathbf{f}, \qquad \mathbf{x} \in \Omega, t \in [0, 2\pi] \tag{1a}$$

$$\nabla \cdot \mathbf{u} = 0, \qquad \mathbf{x} \in \Omega, t \in [0, 2\pi] \tag{1b}$$

where $\mathbf{u} = (u, v, w)$ are the velocity components, aligned with the Cartesian coordinate directions (x, y, z) and $\mathbf{f}$ represents an external force factor. Here Re is the Reynolds number and α is the Womersley number.

We make the particular choice of periodic boundary conditions in time

$$\mathbf{u}(\mathbf{x}, 0) = \mathbf{u}(\mathbf{x}, 2\pi), \quad \mathbf{x} \in \Omega, \tag{2}$$

and Dirichlet boundary conditions in space fixing $\mathbf{u}_{bc} \in C^1(\partial\Omega)$ on the boundary $\partial\Omega$

$$\mathbf{u}(\mathbf{x}, t) = \mathbf{u}_{bc}(\mathbf{x}, t), \quad \mathbf{x} \in \partial\Omega. \tag{3}$$

Thanks to the formulation of Eq. (1a) the N-S equations are dimension-less and we can consider simply $t \in [0, 2\pi)$, that could be scale to any other finite interval in time.

The N-S problem can be conveniently written in an algebraic notation:

$$A\begin{bmatrix}\mathbf{u}\\ p\end{bmatrix} = \begin{pmatrix}\alpha^2\partial_t - \triangle + N[\mathbf{u}] & \mathrm{Re}\nabla\\ \nabla^* & 0\end{pmatrix}\begin{bmatrix}\mathbf{u}\\ p\end{bmatrix} = \begin{bmatrix}\mathbf{f}\\ 0\end{bmatrix} = \mathbf{b}, \tag{4}$$

where $N[\mathbf{u}] = \mathbf{u} \cdot \nabla$ is the nonlinear term. Neglecting N we obtain the Stokes equations, a simpler version of N-S equations. Notice the zero entry for the pressure in A arising from Eq. (1b).

An analytical solution of Eq. (4) is derived in the simplified case of an oscillating pulsatile flow in the x direction with a parabolic profile, where the x velocity depends just on the y coordinate in space, i.e. $\mathbf{u} = (u(y, t), 0, 0)$. We used this solution to validate our implementation and for the error analysis.

3 Discretization

In the following we consider the spatial domain $\Omega = (0, L_x) \times (0, L_y) \times (0, L_z)$ discretized with a uniform Cartesian grid; no error is introduced by approximating the domain. We use a staggered grid in space, where the components of $\mathbf{u} = (u, v, w)$ and p are defined in different nodes (see Fig. 1).

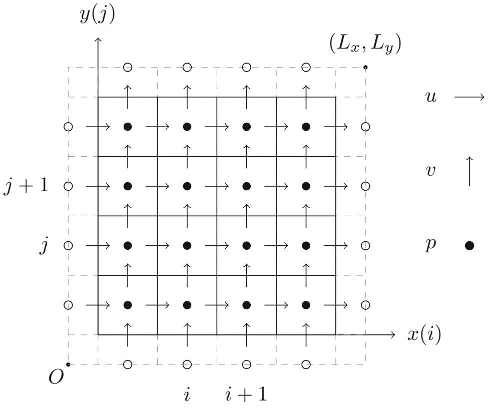


Fig. 1 Staggered grid in a 2D domain and on the boundary. The pressure points are located in i, j nodes (with $i, j \in \mathbb{N}$), u and v are located at $i \pm 1/2, j$ and $i, j \pm 1/2$ nodes. The extension to three spatial dimension is natural: w is defined at $i, j, k \pm 1/2$

The N-S equations will be discretized by means of fourth order finite differences with upwinding for the advection term $N[\mathbf{u}]$.

Because we solve the problem in space-time (without a time stepping technique) we consider the space-time domain $\Omega \times (0, 2\pi)$, which we discretize with $N_x N_y N_z N_T$ grid points,[1] in the form

$$(x_i, y_j, z_k, t_n) = (i\Delta x, j\Delta y, k\Delta z, n\Delta t), \tag{5}$$

where $\Delta x = \frac{L_x}{N_x - 1}$ and $i = 0, \dots, N_x - 1$ (and similarly for y, z and t). Because we consider a uniform grid, we have $\Delta x = \Delta y = \Delta z =: h$. We get the discrete values

$$\begin{aligned} p^n_{ijk} &\approx p(x_i, y_j, z_k, t_n), \quad & u^n_{i\pm1/2,jk} &\approx u(x_{i\pm1/2}, y_j, z_k, t_n), \\ v^n_{ij\pm1/2,k} &\approx v(x_i, y_{j\pm1/2}, z_k, t_n), \quad & w^n_{ijk\pm1/2} &\approx w(x_i, y_j, z_{k\pm1/2}, t_n). \end{aligned}$$

Each of the four variables is defined on a different subset of nodes; if a finite difference stencil requires a variable where it is not defined, it is obtained by linear interpolation.

The momentum equation (1a) is solved in the velocity nodes and the mass equation (1b) in the pressure nodes.

4 Solution Strategy

As first step in the solution process we linearize the momentum equation through a Picard iterative process. In the kth Picard iteration we solve the equations for $\mathbf{u}^k$ and the convective term is transformed:

$$(\mathbf{u} \cdot \nabla)\mathbf{u} \quad \longrightarrow \quad (\mathbf{u}^{k-1} \cdot \nabla)\mathbf{u}^k \tag{6}$$

where we use an initial guess $\mathbf{u}^0$. We iterate until the prescribed tolerance is achieved for the residual.

Successively we apply a multigrid V(ν_1,ν_2) cycle to the linear problem, where ν_1, ν_2 are the pre/post smoothing steps used.

Multigrid methods are a well established solution strategy for elliptic PDEs, but are also applied to parabolic and hyperbolic problems. In particular for elliptic problems multigrid is proven to be an optimal solver, with an h-independent convergence rate.

[1] N_x is the number of points in the x direction and similarly for y, z and time.

We used a fourth order accurate discretization but it is convenient to use a low(er) order discrete operator as smoother in the multigrid cycle; smoothing should be fast but it does not need to be highly accurate. Moreover for higher-order discretizations good smoothing properties are lost (see [6]). Thus, we employ a *defect correction* scheme to obtain high order accuracy from a low order operator. Let's consider two discrete operators: the high order one A_h (4th order in our case) and a lower order one $\hat{A}_h$ (2nd order). The main idea is to solve $A_h\mathbf{x}_h = \mathbf{b}_h$ trough an iterative process where we just evaluate $A_h\mathbf{x}_h$ and solve with $\hat{A}_h$. The nth iteration, with a modified RHS, can be written as

$$\hat{A}_h\mathbf{x}_h^n = \mathbf{b}_h - A_h\mathbf{x}_h^{n-1} + \hat{A}_h\mathbf{x}_h^{n-1}. \tag{7}$$

The defect correction will be performed outside of the multigrid V-cycle.

In the next paragraph we give an outlook on the solution algorithm to solve the non liner system $A_h[\mathbf{x}_h]\mathbf{x}_h = \mathbf{b}_h$; we drop the index h for notational convenience.

Algorithm 1 Solve $A[\mathbf{x}]\mathbf{x} = \mathbf{b}$

```
Initial guess $\mathbf{x}^0$, high order operator $A$, low order operator $\hat{A}$, $k = 0$
Picard iteration process:
while $\left\|\mathbf{b} - A[\mathbf{x}^k]\mathbf{x}^k\right\| > \epsilon$ do
    Linearization of the high order discrete operator: $A^k = A[\mathbf{x}^k]$
    Linearization of the low order discrete operator: $\hat{A}^k = \hat{A}[\mathbf{x}^k]$
    $\mathbf{x}_0^k = \mathbf{x}^k$
    for $n = 0, \dots, N_{\max} - 1$ do
        Compute right hand side for the defect correction:
            $\hat{\mathbf{b}} = \mathbf{b} - A^k\mathbf{x}_n^k + \hat{A}^k\mathbf{x}_n^k$
        Find $\mathbf{x}_{n+1}^k$ solving $\hat{A}^k\mathbf{x}_{n+1}^k = \hat{\mathbf{b}}$ using a multigrid V($\nu_1$,$\nu_2$) cycle and $\mathbf{x}_n^k$ as initial state
    end for
    $\mathbf{x}^{k+1} = \mathbf{x}_{N_{\max}}^k$
    $k = k + 1$
end while
```

where $N_{\max}$ is the number of multigrid cycles performed in one Picard iteration. We employ a V(4,4) cycle with standard transfer operators and a LAPACK optimized GMRES to solve the coarse level system. Of course a key component in the context of a multigrid algorithm is the choice of an effective smoothing method. Because of the pressure zero block in Eq. (4) standard smoothing methods fail to converge. Alternative possibilities are the class of distributive smoothing algorithms (e.g. DGS [7]) or SCGS ("box smoothing"). The latter will be used in this study.

In order to analyze the smoothing properties of SCGS we employed the local Fourier analysis (LFA) as proposed by Brandt [8]. This technique is rigorous only on an infinite grid, but it gives realistic estimates for the convergence of a smoothing scheme on a finite grid, for example see [13]. In fact, with LFA we can compute

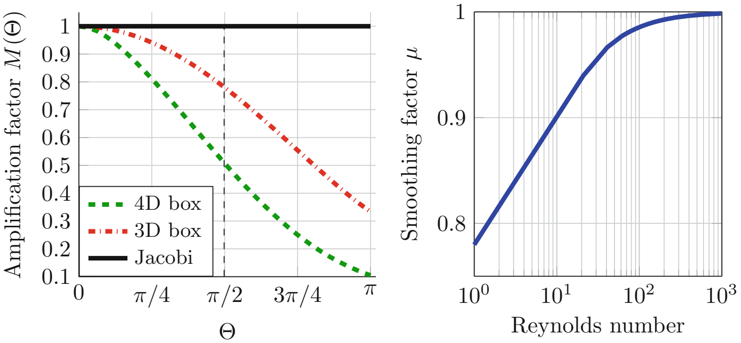


Fig. 2 *Left*: the amplification factor $M(\Theta)$ of SCGS for the whole frequency spectrum for $\mathrm{Re} = \alpha = 1$. $M(\Theta)$ is defined as the spectral radius of the iteration matrix. We compare different box shapes extending in all 4D or just in 3D. The "Jacobi boxes" are decoupled, i.e. non intersecting each other. *Right*: we show how μ depends on the Reynolds number; the smoothing factor μ is defined as $\mu = \max\{|M(\Theta)| : \pi/2 \le \Theta \le \pi\}$

the smoothing factor (usually denoted by μ), a measure of the reduction of high frequency error components per smoothing step. We refer to [9] and [10] for the steady Stokes equations, discretized with a standard grid in 2D. We extend this analysis to the 3D unsteady Stokes equations, with boxes in space and space-time on a staggered grid. In Fig. 2 we show some results of LFA. For more details regarding this analysis and the solution strategy in general we refer to [1].

5 Numerical Experiments

Our implementation is based on the library TRILINOS [11] and the framework P-IMPACT, developed at ETH Zürich, an extension of the library IMPACT [12] designed to solve CFD problems with finite differences in parallel. This library allows to decompose the space-time domain on a Cartesian MPI processor grid and exchange data between neighboring processors trough ghost cell layers. All the measurements are obtained on the Euler[2] cluster. If not stated differently, we used $\mathrm{Re} = \alpha = 1$ and $N \sim 3.2 \cdot 10^5$ unknowns equally distributed in the four dimensions. We define the scaled residual as $d = \|\mathbf{d}_h\|_2 / \sqrt{N}$, with $\mathbf{d}_h = \mathbf{b}_h - A_h \mathbf{x}_h$. In Fig. 3 we show the smoothing effect of SCGS, on the N-S equations considering the effect of alternating the direction in the sweep and introducing a under-relaxation parameter ω for the update in case of $\mathrm{Re} = 100$. In Fig. 4 we show the convergence of the multigrid algorithm and of the outer Picard iterative process, i.e. the overall convergence. In Fig. 5 weak and strong scaling of the parallel solver are shown.

[2] http://brutuswiki.ethz.ch/brutus/Getting_started_with_Euler

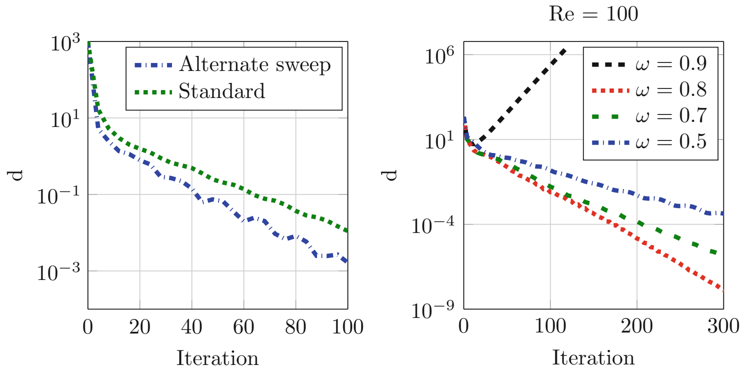


Fig. 3 *Left*: the beneficial effect of alternating the sweep directions. *Right*: for higher Reynolds number underelaxation is necessary. We can see that the convergence of the smoother is initially fast but it degrades after ∼10 iterations

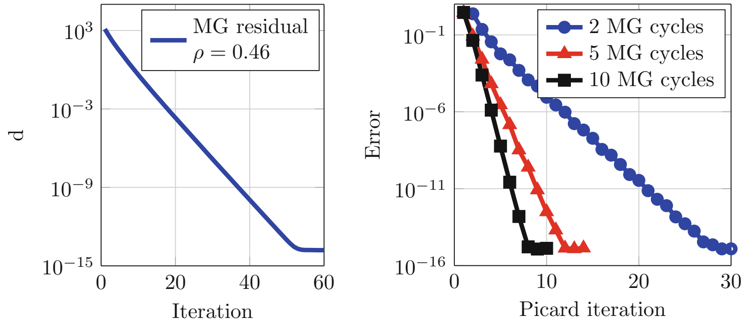


Fig. 4 *Left*: multigrid convergence, ρ is the convergence factor. *Right*: convergence of Picard iterations with different numbers of multigrid cycles per iteration

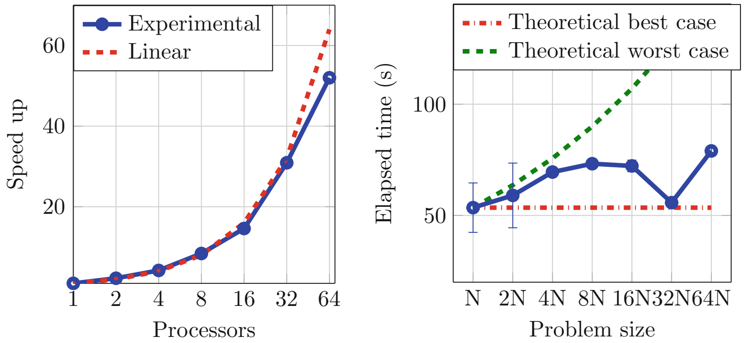


Fig. 5 *Left*: strong scaling. *Right*: weak scaling

6 Conclusions

In this work we have investigated and built a parallel space-time solver for the Navier–Stokes equations in $3 + 1$ dimensions with periodic boundary conditions. In particular, the case of a pulsatile flow is analyzed and simulated by means of high order finite difference on a staggered grid.

The SCGS smoothing process, necessary for the multigrid algorithm, has been studied for the simplified Stokes case. In the context of the local Fourier analysis, we compared this method for different boxes. A box in four dimensions turned out to be the one with the best smoothing factor, both in the LFA theory and in the numerical experiments. For "high" Reynolds numbers under-relaxation may be necessary for SCGS to be convergent.

We have used a space-time multigrid algorithm with a defect correction technique to accelerate the convergence of SCGS, obtaining a convergence factor $\rho \simeq 0.46$, stable to machine precision and h-independent. We achieved convergence to machine precision with just ten Picard steps, in the best case. For Re $\gtrsim$ 100 the multigrid algorithm is unstable for the studied problem and a more thorough analysis should be carried out. Stabilization techniques may be necessary as the introduction of artificial diffusion in the problem.

Finally we have analyzed the scalability of the software subdividing the domain between multiple processors in all four dimensions. We obtained a $\sim 50\times$ times speedup using 64 processors, and high efficiency especially for 32 or lower numbers of processors.

Acknowledgements Pietro Benedusi and Rolf Krause would like to thank the Swiss National Science Foundation (SNF) and the Deutsche Forschungsgemeinschaft for supporting this work in the framework of the project "**ExaSolvers – Extreme Scale Solvers for Coupled Systems**", SNF project number 145271, within the DFG-Priority Research Program 1684 "SPPEXA- Software for Exascale Computing"

References

1. P. Benedusi, A parallel multigrid solver for time-periodic incompressible Navier–Stokes equations. Master thesis, USI Lugano, ICS (2015)
2. S.P. Vanka, Block-implicit multigrid solution of Navier–Stokes equations in primitive variables. J. Comput. Phys. **65**, 138–156 (1986)
3. I.V. Pivkin, P.D. Richardson, D.H. Laidlaw, G.E. Karniadakis, Combined effects of pulsatile flow and dynamic curvature on wall shear stress in a coronary artery bifurcation model. J. Biomech. **38**, 1283–1290 (2015)
4. M. Mehrabi, S. Setayeshi, Computational fluid dynamics analysis of pulsatile blood flow behavior in modelled stenosed vessels with different severities. Math. Probl. Eng. **2012**, 13 (2012)
5. J.B. Grotberg, Pulmonary flow and transport phenomena. Annu. Rev. Fluid Mech. **26**, 529–571 (1994)

6. B. Koren, Multigrid and defect correction for the steady Navier–Stokes equations. J. Comput. Phys. **87**, 25–46 (1989)
7. W. Ming, C. Long, Multigrid methods for the stokes equations using distributive Gauss–Seidel relaxations based on the least squares commutator. J. Sci. Comput. **56** (2013)
8. A. Brandt, Multi-level adaptive solutions to boundary-value problems. Math. Comput. **31**, 333–390 (1977)
9. L.B. Zhang, Box-line relaxation schemes for solving the steady incompressible Navier–Sotkes equaitons using second order upwind differincing. J. Comput. Math. **13**, 32–39 (1991)
10. S. Sivaloganathan, The use of local mode analysis in the design and comparison of multigrid methods. Comput. Phys. Comm. **65**, 246–252 (1991)
11. Heroux et al., An overview of the trilinos project. ACM Trans. Math. Softw. **31**, 397–423 (2005)
12. R. Henniger, D. Obrist, L. Kleiser, High-order accurate solution of the incompressible Navier–Stokes equations on massively parallel computers. J. Comput. Phys. **229**, 3543–3572 (2010)
13. J. Linden et al., Multigrid for the steady-state incompressible Navier–Stokes equations: a survey, in *11th International Conference on Numerical Methods in Fluid Dynamics* (Springer, Berlin/Heidelberg, 1989)

Discretization and Parallel Iterative Schemes for Advection-Diffusion-Reaction Problems

Abdullah Ali Sivas, Murat Manguoğlu, J.H.M. ten Thije Boonkkamp, and M.J.H. Anthonissen

Abstract Conservation laws of advection-diffusion-reaction (ADR) type are ubiquitous in continuum physics. In this paper we outline discretization of these problems and iterative schemes for the resulting linear system. For discretization we use the finite volume method in combination with the complete flux scheme. The numerical flux is the superposition of a homogeneous flux, corresponding to the advection-diffusion operator, and the inhomogeneous flux, taking into account the effect of the source term (ten Thije Boonkkamp and Anthonissen, J Sci Comput 46(1):47–70, 2011). For a three-dimensional conservation law this results in a 27-point coupling for the unknown as well as the source term. Direct solution of the sparse linear systems that arise in 3D ADR problems is not feasible due to fill-in. Iterative solution of such linear systems remains to be the only efficient alternative which requires less memory and shorter time to solution compared to direct solvers. Iterative solvers require a preconditioner to reduce the number of iterations. We present results using several different preconditioning techniques and study their effectiveness.

1 Discretization and Iterative Solution

We consider a stationary conservation law of advection-diffusion-reaction type, viz.

$$\nabla \cdot (\mathbf{u}\varphi - \varepsilon \nabla \varphi) = s, \tag{1}$$

A.A. Sivas (✉)
Institute of Applied Mathematics, METU, 06800, Ankara, Turkey
e-mail: sivas@metu.edu.tr

M. Manguoğlu
Department of Computer Engineering, Institute of Applied Mathematics, METU, 06800, Ankara, Turkey
e-mail: manguoglu@ceng.metu.edu.tr

J.H.M. ten Thije Boonkkamp • M.J.H. Anthonissen
Department of Mathematics and Computer Science, TUE, Eindhoven, The Netherlands
e-mail: j.h.m.tenthijeboonkkamp@tue.nl; m.j.h.anthonissen@tue.nl

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_27

where $\mathbf{u} = u\,\mathbf{e}_x + v\,\mathbf{e}_y + w\,\mathbf{e}_z$ is a mass flux or (drift) velocity, $\varepsilon \geq \varepsilon_{\min} > 0$ a diffusion coefficient, and s a source term describing, e.g., chemical reactions or ionization. The unknown φ is then the mass fraction of one of the constituent species in a chemically reacting flow or a plasma. The parameters ε and s are usually (complicated) functions of φ whereas the vector field $\mathbf{u}$ has to be computed from (flow) equations corresponding to (1). However, for the sake of discretization, we will consider these parameters as given functions of the spatial coordinates.

Associated with equation (1) we introduce the flux vector $\mathbf{f}$, defined by $\mathbf{f} := \mathbf{u}\varphi - \varepsilon\nabla\varphi$. Consequently, equation (1) can be concisely written as $\nabla \cdot \mathbf{f} = s$. Integrating this equation over a fixed domain Ω and applying Gauss's theorem we obtain the integral form of the conservation law, i.e.,

$$\oint_{\Gamma} (\mathbf{f}, \mathbf{n})\, dS = \int_{\Omega} s\, dV, \tag{2}$$

where $\mathbf{n}$ is the outward unit normal on the boundary $\Gamma = \partial\Omega$. In the FVM [4] we cover the domain with a finite number of disjunct control volumes or cells and impose the integral form (2) for each of these cells.

In three-dimensional Cartesian coordinates, we first choose grid points $\mathbf{x}_{i,j,k} = (x_i, y_j, z_k)$ where the unknown φ has to be approximated. Next, we choose control volumes $\Omega_{i,j,k} := (x_{i-1/2}, x_{i+1/2}) \times (y_{j-1/2}, y_{j+1/2}) \times (z_{k-1/2}, z_{k+1/2})$. Here $x_{i\pm 1/2} := \frac{1}{2}(x_i + x_{i\pm 1})$ etc. The boundary of control volume $\Omega_{i,j,k}$ is the union of six surfaces $\Gamma_{i\pm 1/2,j,k}$, $\Gamma_{i,j\pm 1/2,k}$ and $\Gamma_{i,j,k\pm 1/2}$. Taking $\Omega = \Omega_{i,j,k}$ in conservation law (2) and approximating all integrals with the midpoint rule, we find

$$\begin{aligned}&\left(f_{x,i+1/2,j,k} - f_{x,i-1/2,j,k}\right) \Delta y\, \Delta z\\ &\qquad + \left(f_{y,i,j+1/2,k} - f_{y,i,j-1/2,k}\right) \Delta x\, \Delta z\\ &\qquad + \left(f_{z,i,j,k+1/2} - f_{z,i,j,k-1/2}\right) \Delta x\, \Delta y \doteq s_{i,j,k}\, \Delta x\, \Delta y\, \Delta z,\end{aligned} \tag{3}$$

where we have used that $\mathbf{f} = f_x\,\mathbf{e}_x + f_y\,\mathbf{e}_y + f_z\,\mathbf{e}_z$. The FVM has to be completed with numerical approximations F_x, F_y and F_z for the fluxes f_x, f_y and f_z in (3).

We first derive the x-component of the flux, $f_{x,i+1/2,j,k}$, by solving a local one-dimensional problem. Consider the flux $f := u\varphi - \varepsilon\frac{d\varphi}{dx}$ and the model BVP:

$$\frac{d}{dx}(f) = s, \quad x_i < x < x_{i+1}, \qquad \varphi(x_i) = \varphi_i, \quad \varphi(x_{i+1}) = \varphi_{i+1}. \tag{4}$$

We assume $\varepsilon > 0$ and s to be sufficiently smooth functions of x. Solving (4), we find $f_{i+\frac{1}{2}} = f^{\mathrm{hom}}_{i+\frac{1}{2}} + f^{\mathrm{inh}}_{i+\frac{1}{2}}$, see [8] for details, where we have introduced the homogeneous flux f^{hom} and the inhomogeneous flux f^{inh}. These fluxes correspond to the advection-diffusion operator and the source term, respectively. If both u and ε are constant, we can write the homogeneous flux as

$$f^{\mathrm{hom}}_{i+\frac{1}{2}} = \frac{\varepsilon}{\Delta x}\Big(B(-P)\varphi_i - B(P)\varphi_{i+1}\Big), \tag{5}$$

in which $B(x) := x/(\mathrm{e}^x - 1)$ and the Péclet number $P := \frac{u}{\varepsilon}\,\Delta x$. For the inhomogeneous flux we find

$$f^{\mathrm{inh}}_{i+\frac{1}{2}} = \int_0^1 G(\sigma;P)\, s(x(\sigma))\, d\sigma, \tag{6}$$

where we have used the *normalized coordinate* $\sigma(x) := (x - x_i)/\Delta x$ and the *Green's function for the flux*

$$G(\sigma;P) := \begin{cases} \dfrac{\mathrm{e}^{-\sigma P} - 1}{\mathrm{e}^{-P} - 1}, & \text{for} \quad 0 \le \sigma \le 1/2, \\[2ex] -\dfrac{\mathrm{e}^{(1-\sigma)P} - 1}{\mathrm{e}^{P} - 1}, & \text{for} \quad 1/2 < \sigma \le 1. \end{cases} \tag{7}$$

For the numerical fluxes, two averages are used: the normal *arithmetic* average $\overline{\varepsilon}_{i+\frac{1}{2}} := (\varepsilon_i + \varepsilon_{i+1})/2$ and a *weighted* average $\widetilde{\varepsilon}_{i+\frac{1}{2}} := W(-\overline{P}_{i+\frac{1}{2}})\varepsilon_i + W(\overline{P}_{i+\frac{1}{2}})\varepsilon_{i+1}$. The weight function used here is $W(x) := (\mathrm{e}^x - 1 - x)/(x\,(\mathrm{e}^x - 1))$. We use (5) to find the following *numerical homogeneous flux*

$$F^{\mathrm{hom}}_{i+\frac{1}{2}} = \alpha_{i+\frac{1}{2}}\varphi_i - \beta_{i+\frac{1}{2}}\varphi_{i+1}, \tag{8a}$$

$$\alpha_{i+\frac{1}{2}} := B\big(-\overline{P}_{i+\frac{1}{2}}\big)\frac{\widetilde{P}_{i+\frac{1}{2}}}{\overline{P}_{i+\frac{1}{2}}}\frac{\widetilde{\varepsilon}_{i+\frac{1}{2}}}{\Delta x}, \quad \beta_{i+\frac{1}{2}} := B\big(\overline{P}_{i+\frac{1}{2}}\big)\frac{\widetilde{P}_{i+\frac{1}{2}}}{\overline{P}_{i+\frac{1}{2}}}\frac{\widetilde{\varepsilon}_{i+\frac{1}{2}}}{\Delta x}. \tag{8b}$$

The *numerical inhomogeneous flux* is based on (6). We take $s(x)$ equal to s_i on the interval $(0, 1/2)$ and equal to s_{i+1} on $(1/2, 1)$. Next we integrate the Green's function to find

$$F^{\mathrm{inh}}_{i+\frac{1}{2}} := \gamma_{i+\frac{1}{2}} s_i - \delta_{i+\frac{1}{2}} s_{i+1}, \tag{9a}$$

$$\gamma_{i+\frac{1}{2}} := C(-\overline{P}_{i+\frac{1}{2}})\,\Delta x, \quad \delta_{i+\frac{1}{2}} := C(\overline{P}_{i+\frac{1}{2}})\,\Delta x, \tag{9b}$$

in which $C(x) := (\mathrm{e}^{\frac{1}{2}x} - 1 - \frac{1}{2}x)/(x\,(\mathrm{e}^x - 1))$. Note: $C(x) \to 1/8$ for $x \to 0$, $C(x) \to 0$ for $x \to \infty$, and $C(x) \to 1/2$ for $x \to -\infty$. Hence for small Péclet numbers, the coefficients γ and δ will be approximately equal and the inhomogeneous flux is small. For large (positive or negative) Péclet numbers, the upwind value of s has a dominant contribution. Adding (8) and (9), we obtain the following *numerical complete flux*:

$$F_{i+\frac{1}{2}} = \alpha_{i+\frac{1}{2}}\varphi_i - \beta_{i+\frac{1}{2}}\varphi_{i+1} + \gamma_{i+\frac{1}{2}}s_i - \delta_{i+\frac{1}{2}}s_{i+1}. \tag{10}$$

We will now combine the one-dimensional schemes to derive a numerical scheme for the 3D equation (1). The key idea is to include the cross-fluxes $\partial f_y/\partial y$ and $\partial f_z/\partial z$ in the evaluation of the flux in x-direction. This reduces the crosswind diffusion and leads to much sharper (interior) layers for advection-dominated flows. In [8] we have shown that for 2D problems, the inclusion of cross-flux terms is essential to maintain second order accuracy, whereas the homogeneous flux scheme (without cross-fluxes) reduces to first order. Hence, we determine the numerical flux $F_{x,i+\frac{1}{2},j,k}$ from the quasi-one-dimensional boundary value problem:

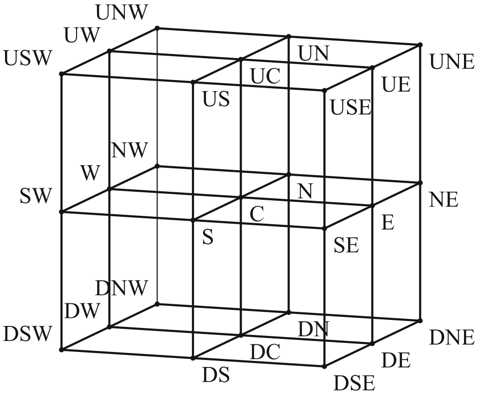
$$\frac{\partial}{\partial x}\left(\left(u\varphi - \varepsilon\frac{\partial\varphi}{\partial x}\right)\right) = s_x, \qquad x_i < x < x_{i+1},\ y = y_j,\ z = z_k, \tag{11a}$$

$$\varphi(\mathbf{x}_{i,j,k}) = \varphi_{i,j,k}, \qquad \varphi(\mathbf{x}_{i+1,j,k}) = \varphi_{i+1,j,k}, \tag{11b}$$

where the modified source term s_x is defined by $s_x := \alpha s - \beta(\partial f_y/\partial y + \partial f_z/\partial z)$, with α and β coefficients that are yet to be determined. If we take $\beta = 1$, the cross-fluxes are completely included; taking $\beta = 0$ ignores them.

The derivation of the expression for the numerical flux is essentially the same as for (10), the main difference being the inclusion of the cross-fluxes $\partial f_y/\partial y$ and $\partial f_z/\partial z$ in the source term. In the computation of s_x we replace $\partial f_y/\partial y$ by its central difference approximation and for f_y we take the homogeneous numerical flux. We treat $\partial f_z/\partial z$ in the same way. Similar procedures apply to the y- and z-components of the flux. We shall take $\beta = 1/2$ in the numerical simulations. Adding the three one-dimensional problems in x, y and z-direction, we find that we need to choose $\alpha = (1 + 2\beta)/3$ for consistency. Substitution of the numerical fluxes presented above into (3) leads to a 27-point stencil for the unknown φ. The points of the stencil are presented in Fig. 1. We denote the resulting linear system by $\mathbf{Ax} = \mathbf{b}$. The matrix $\mathbf{A}$ has in general 27 nonzero diagonals.

Fig. 1 Compass notation for points in discretization stencil



The resulting sparse linear system from the 3D discretization is most suitable for iterative solvers since direct solvers are known to scale poorly and memory requirements are usually very high due to fill-in. We use Bi-Conjugate Gradient Stabilized (BiCGStab) with preconditioning. In this paper, we will study and compare the performance of two parallel preconditioners that are based on the sparse approximate inverse (SAI) and incomplete LU (ILU) factorization.

Given a linear system $\mathbf{Ax} = \mathbf{b}$, the sparse approximate inverse idea is to find a sparse matrix $\mathbf{M}$ such that the Frobenius norm of the error $||\mathbf{MA}-\mathbf{I}||_F < \epsilon$ for some $\epsilon > 0$ under a sparsity constraint on $\mathbf{M}$. If the structure of $\mathbf{A}^{-1}$ is known, it can be used for the sparsity pattern of $\mathbf{M}$. If it is not known, one practical approach is to assume that $\mathbf{M}$ has the same structure as $\mathbf{A}^k$ for some k, but as k gets larger this is costly, and there is a limit to ϵ as the structure is fixed. Alternatively, for a given ϵ, trying to find a structure for $\mathbf{M}$ that satisfies given ϵ is another possibility.

SAI type of preconditioners are expected to be scalable on parallel computing platforms since computing the preconditioner matrix can be split into completely independent linear least squares problems. Another reason is that applying the preconditioner is just a matrix-vector multiplication which is usually possible to parallelize.

The main idea in ILU-type preconditioners is to find an approximate LU factorization of the coefficient matrix $\mathbf{A}$ where $\mathbf{M} = \widetilde{\mathbf{L}}\widetilde{\mathbf{U}} \approx \mathbf{A}$ which is sparser than complete LU factorization and $\mathbf{M}^{-1}\mathbf{A}$ has a more favorable eigenvalue distribution. This is either achieved by assuming a structure beforehand or by dropping elements which are less than some threshold in absolute value. Parallel ILU types of preconditioners are less amenable to parallelism. This is because of the inherently sequential nature and limited parallelism of the triangular solves as well as the incomplete factorization processes. We refer the reader to surveys [1] and [2] for a detailed discussion on various other preconditioning techniques.

2 Numerical Results

2.1 A Three-Dimensional Flow Problem

We consider the following flow problem. It is a three-dimensional extension of [8, Section 8, Example 3]. The problem domain is given by $-1 \leq x \leq 1$, $0 \leq y \leq 1$ and $0 \leq z \leq 1$. The unknown φ satisfies the following partial differential equation

$$\nabla \cdot (\mathbf{u}\varphi - \varepsilon\nabla\varphi) = 0, \qquad \text{in } (-1,1)\times(0,1)\times(0,1) \tag{12}$$

with velocity field $\mathbf{u}(x,y,z) = (1-x^2)y(1-2z)\,\mathbf{e}_x + x(1-y^2)(1-2z)\,\mathbf{e}_y + 4xyz(1-z)\,\mathbf{e}_z$, see Fig. 2. We impose the following boundary conditions (c is a constant that determines the steepness of the profile; we take $c = 10$)

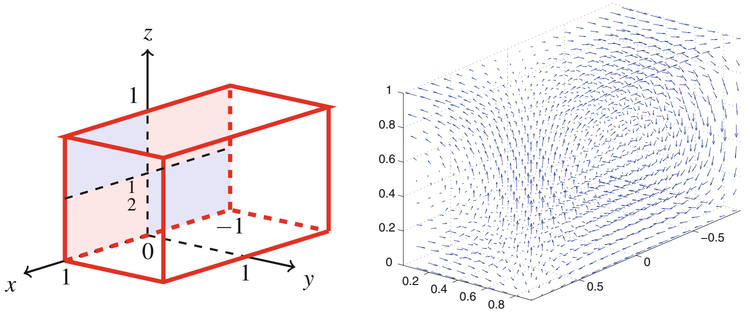


Fig. 2 The problem domain and velocity field for the flow problem. The inlet is *lower left* and *upper right rectangle* the outlet *lower right rectangle* and *upper left rectangle*

- At the inlet (given by $y = 0, x \geq 0, 0 \leq z \leq 1/2$ and $y = 0, x \leq 0, 1/2 \leq z \leq 1$) we set $\varphi(x, y, z) = \big(1 + \tanh(c(2x + 1))\big)z$
- At the outlet (given by $y = 0, x < 0, 0 \leq z < 1/2$ and $y = 0, x > 0, 1/2 < z \leq 1$) we set $\frac{\partial \varphi}{\partial y}(x, y, z) = 0$
- At the front ($x = 1$), back ($x = -1$), right ($y = 1$), bottom ($z = 0$), and top ($z = 1$), we set $\varphi(x, y, z) = \big(1 + \tanh(c(2x + 1))\big)(1 - y)z$.

2.2 *Parallel Scalability*

All the runs are performed on a single shared memory node which has two 10-core Intel Xeon E5-2650v3 2.3 GHz processors (total 20 cores) and 64GB of memory, running on CentOS 6.6 operating system.

As the iterative solver environment we use Hypre [5] version `2.10.0b`. Hypre provides a parallel environment for various iterative solvers and preconditioners using MPI. In the following runs we use preconditioned BiCGStab [9] as the Krylov iterative solver. Two preconditioners we have experimented with are: Euclid [6] and ParaSails [3]. Euclid is a parallel ILU implementation and ParaSails is a parallel SAI type of preconditioner. Euclid has support for Parallel-ILU(k) and Block Jacobi ILU(k) in its current implementation,we have used Parallel-ILU(1). For ParaSails we set parameters so that the preconditioner has the same sparsity as the coefficient matrix in the worst case. We also set a threshold value of 0.2 when determining the sparsity structure of **M**, based on the sparsity structure of **A** by ignoring smaller nonzeros in absolute value based on the threshold. The filter value is set to be 0.05 which drops smaller elements in absolute value from the preconditioner after they are computed.

We use Pardiso [7] direct solver which is a part of Intel MKL version `11.2.1` to compare against. We used default parameters of Pardiso solver.

Table 1 Number of iterations for the iterative solvers using different diffusion (ε) coefficients

	$\varepsilon = 1$		$\varepsilon = 10^{-5}$	
Cores	ParaSails	Euclid	ParaSails	Euclid
1	226	44	1178	94
2	212	44	1183	125
4	215	49	1186	164
8	236	45	1191	235
16	220	51	1151	415
20	213	52	1177	–

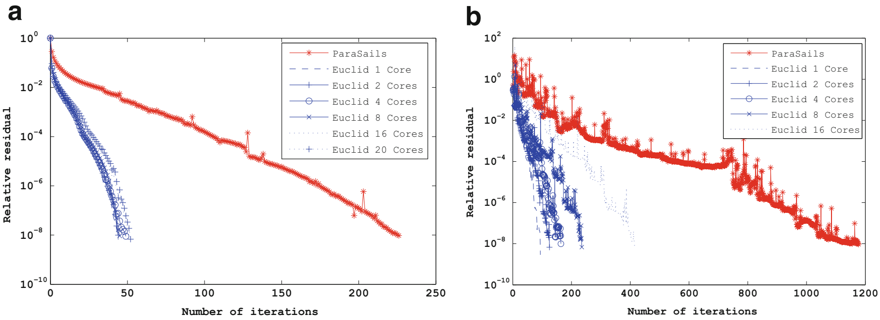


Fig. 3 Residual history. (**a**) $\varepsilon = 1$. (**b**) $\varepsilon = 10^{-5}$

We have experimented with two linear systems of the same size but with different diffusion coefficients of 1 and 10^{-5}. Initially the systems are generated including the boundary conditions but for all linear solvers we remove the Dirichlet boundary conditions. After removing the boundary conditions, the resulting systems have 992,319 unknowns and 18,467,751 and 17,452,253 nonzeros, respectively for diffusion coefficients (ε) of 1 and 10^{-5}. Smaller ε results in a sparser coefficient matrix. For the iterative solvers, stopping criteria is L_2-norm of the relative residual to be less than or equal to 10^{-8} or maximum number of iterations (10,000) to be reached. We measure the elapsed wall clock time and the number of iterations.Each MPI process is mapped on a single core.

Table 1 shows the number of iterations of ParaSails and Euclid for different diffusion coefficients. Although ParaSails requires a larger number of iterations, its iteration count does not depend on the number of cores (and the number of processes) while the required number of iterations for Euclid increases with increasing number of cores, especially for smaller ε. In Fig. 3a, b, convergence histories of Euclid and ParaSails are given for $\varepsilon = 1$ and 10^{-5}, respectively. We also note that for $\varepsilon = 10^{-5}$ and 20 cores, using BiCGStab with Euclid preconditioner has failed before it reaches the maximum number of iterations.

In Fig. 4a, b, the required time to solution for $\varepsilon = 1$ and $\varepsilon = 10^{-5}$ are given, respectively, using Pardiso, Euclid and ParaSails. As expected the direct solver is the slowest due to fill-in. When $\varepsilon = 10^{-5}$, ParaSails is better than Euclid for large

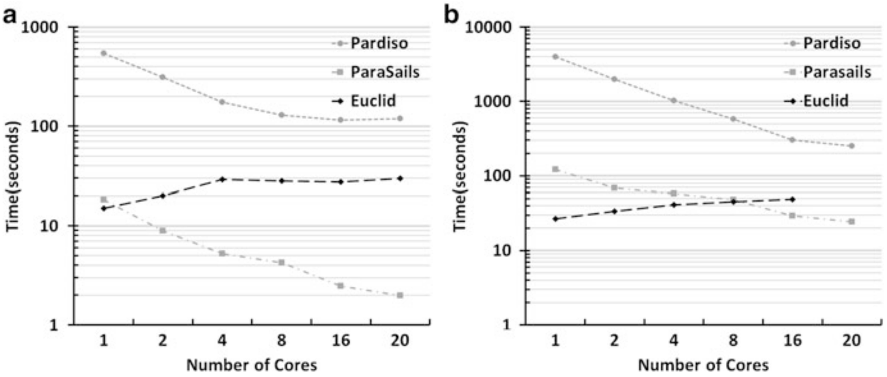


Fig. 4 Parallel running time in seconds. (**a**) $\varepsilon = 1$. (**b**) $\varepsilon = 10^{-5}$

number of cores. For $\varepsilon = 1$, on the other hand, ParaSails is the faster and scales much better than Euclid throughout. This is probably due to Euclid's increasing number of iterations as we increase the number of cores and due to the inherently sequential nature of triangular solves.

3 Conclusions

We have presented a new 3D formulation and discretization of advection-diffusion-reaction equations using the complete flux scheme. Resulting linear systems are solved directly using Pardiso and iteratively using ILU and SAI type of preconditioners on a parallel platform. Results show that for various diffusion coefficients, SAI preconditioned BiCGStab is the more scalable and requires less time to solution compared to the ILU preconditioned BiCGStab and to the direct solver.

Acknowledgements This work was partially supported by Turkish Academy of Sciences Distinguished Young Scientist Award M.M./TUBA-GEBIP/2012–19. Authors would like to thank Bülent Karasözen for his suggestions during the implementation.

References

1. O. Axelsson, A survey of preconditioned iterative methods for linear systems of algebraic equations. BIT Numer. Math. **25**(1), 165–187 (1985)
2. M. Benzi, Preconditioning techniques for large linear systems: a survey. J. Comput. Phys. **182**(2), 418–477 (2002)
3. E. Chow, Parallel implementation and practical use of sparse approximate inverse preconditioners with a priori sparsity patterns. Int. J. High Perform. Comput. Appl. **15**(1), 56–74 (2001)

4. R. Eymard, T. Gallouët, R. Herbin, Finite volume methods, in *Handbook of Numerical Analysis*, ed. by P.G. Ciarlet, J.L. Lions, vol. VII (Elsevier, North-Holland, 2000), pp. 713–1020
5. R. Falgout, U. Yang, Hypre: a library of high performance preconditioners. Comput. Sci. ICCS **2002**, 632–641 (2002)
6. D. Hysom, A. Pothen, A scalable parallel algorithm for incomplete factor preconditioning. SIAM J. Sci. Comput. **22**(6), 2194–2215 (2001)
7. O. Schenk, K. Gärtner, Solving unsymmetric sparse systems of linear equations with PARDISO. Future Gener. Comput. Syst. **20**(3), 475–487 (2004)
8. J.H.M. ten Thije Boonkkamp, M.J.H. Anthonissen, The finite volume-complete flux scheme for advection-diffusion-reaction equations. J. Sci. Comput. **46**(1), 47–70 (2011)
9. H.A. van der Vorst, Bi-CGSTAB: a fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems. SIAM J. Sci. Stat. Comput. **13**(2), 631–644 (1992)

A Simple Proposal for Parallel Computation Over Time of an Evolutionary Process with Implicit Time Stepping

Eleanor McDonald and Andy Wathen

Abstract Evolutionary processes arise in many areas of applied mathematics, however since the solution at any time depends on the solution at previous time steps, these types of problems are inherently difficult to parallelize. In this paper, we make a simple proposal of a parallel approach for the solution of evolutionary problems with implicit time step schemes. We derive and demonstrate our approach for both the linear diffusion equation and the convection-diffusion equation. Using an all-at-once approach, we solve for all time steps simultaneously using a parallelizable over time preconditioner within a standard iterative method.

1 Introduction

Evolutionary processes have been extensively studied for many years and as we approach the limits of computational speeds on a single processor, the use of massively parallel computer architectures is seen as the way forward. The inherent problem with solving evolutionary problems on such systems is the causality principle; the solution at each time step depends on the solution at previous steps and this presents significant difficulty to parallel computations over time. Although explicit time stepping schemes are readily parallelizable, this places significant restriction on the time step size, τ, in order to ensure stability. Therefore methods which can be applied to unconditionally stable implicit methods allow parallelization with larger time step sizes.

Significant research has been conducted on time domain parallelization and a recent comprehensive review can be found in [6]. As described in this review, methods can generally be classified into either multiple shooting methods (such as the "parareal" method [10] or PFASST [4]), domain decomposition and waveform relaxation methods such as in [5], space-time multigrid methods [8, 11], or direct methods [7]. Our proposal falls into none of these categories as it contains no coars-

E. McDonald (✉) • A. Wathen
Mathematical Institute, University of Oxford, Woodstock Road, Oxford, OX2 6GG, UK
e-mail: mcdonalde@maths.ox.ac.uk; wathen@maths.ox.ac.uk

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_28

ening of the time operator and is based solely on a block diagonal preconditioner to the all-at-once system.

The method we propose is a parallelizable block diagonal preconditioner applied within an iterative method such as GMRES [13] or BiCGSTAB [14]. The solution is computed using the all-at-once method, whereby the solution at all time steps is computed simultaneously. Parallelization of an all-at-once formulation is also presented in [11] where a space-time multigrid approach is used however in our more elementary approach, the block diagonal preconditioner is approximated with multigrid but there is no coarsening in time. Despite having no coarse time propagator, our multigrid approximated preconditioner performs similarly to a block Jacobi preconditioner applied exactly which has provable convergence properties. Thus we are able to take advantage of the multigrid approximation to propose a cheaply applied, parallelizable, and easy to implement preconditioner.

In this short paper, we will motivate our method firstly on the linear diffusion equation and provide analysis to support the convergence of the method. We will extend this methodology to the convection-diffusion equation and finally present results to illustrate the performance for the model problems considered.

2 Proposal Outline

In order to describe our method, we will begin by considering the solution of the linear diffusion (or heat) equation initial-boundary value problem,

$$\begin{aligned} u_t &= \Delta u + f && \text{in } \Omega \times (0,T], \quad \Omega \subset \mathbb{R}^2 \,\text{or}\, \mathbb{R}^3, \\ u &= g && \text{on } \partial\Omega, \\ u(x,0) &= u_0(x) && \text{at } t = 0. \end{aligned} \tag{1}$$

We will use a finite element discretization in space on a uniform square grid with mesh size h though there is no reason to believe that this is necessary for our proposal. A θ-method will be used to discretize in time with N time steps of size τ_k at the k-th step where $\sum_{k=1}^{N} \tau_k = T$. We note that for $\frac{1}{2} \leq \theta \leq 1$ we have an unconditionally stable implicit scheme. While this method is valid for all θ-schemes we will focus on results for the Backward Euler ($\theta = 1$) and Crank-Nicolson ($\theta = \frac{1}{2}$) schemes.

For general θ-schemes the discretization of (1) gives,

$$M\frac{\mathbf{u}_k - \mathbf{u}_{k-1}}{\tau_k} + K\left(\theta \mathbf{u}_k + (1-\theta)\mathbf{u}_{k-1}\right) = \mathbf{f}_k, \tag{2}$$

for $k = 1, \ldots, N$ where $M \in \mathbb{R}^{n \times n}$ is the standard finite element mass matrix, $K \in \mathbb{R}^{n \times n}$ is the stiffness matrix (the discrete Laplacian), and n is the number of spatial degrees of freedom.

In order to solve this system, the classical approach is to solve the N separate $n \times n$ systems sequentially for $k = 1, 2, \ldots, N$. For large n, an iterative method such as Algebraic Multigrid (AMG) may be used to complete each of these solves. This approach is inherently sequential, hence the method is difficult to parallelize over time. If we regard 1 AMG V-cycle as the main unit of work for each solve, then if we require r V-cycles to solve the $n \times n$ linear system at each time step to the desired accuracy, we require rN V-cycles in total to achieve an accurate solution at all time steps.

2.1 Proposed Approach

Our simple proposal solves all time steps simultaneously using an 'all-at-once' approach. Conceptually, we construct the following linear system which defines the solution at all time steps in one large equation system, $\mathscr{A}\mathbf{u} = \mathbf{f}$, where

$$\mathscr{A} := \begin{bmatrix} M + \tau_1\theta K & & & \\ -M + \tau_2(1-\theta)K & M + \tau_2\theta K & & \\ & \ddots & \ddots & \\ & & -M + \tau_N(1-\theta)K & M + \tau_N\theta K \end{bmatrix}, \qquad (3)$$

and $\mathbf{u} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_N]^T$ and $\mathbf{f} = [(M - \tau_1(1-\theta)K)\mathbf{u}_0 + \tau_1\mathbf{f}_1, \tau_2\mathbf{f}_2, \ldots, \tau_N\mathbf{f}_N]^T$.

We note that the resulting linear system is a huge $nN \times nN$ system however, matrix vector products with $\mathscr{A}$ require only vector products with M and K, thus can be computed simply without actual construction of $\mathscr{A}$. The classical approach would correspond to solution of this system by block forward substitution. In order to solve the system we propose to use a standard preconditioned iterative method such as GMRES or BiCGSTAB; the key consideration is preconditioning.

The preconditioner we propose is an approximation to the block diagonal of $\mathscr{A}$ in (3) and is given by,

$$\mathscr{P}_{MG}^{-1} := \begin{bmatrix} (M + \tau_1\theta K)_{MG} & & & \\ & (M + \tau_2\theta K)_{MG} & & \\ & & \ddots & \\ & & & (M + \tau_N\theta K)_{MG} \end{bmatrix}. \qquad (4)$$

where $(M + \tau_k\theta K)_{MG}$ denotes the application of a single multigrid V-cycle to the matrix $M + \tau_k\theta K \in \mathbb{R}^{n\times n}$. Due to the block diagonal structure, this preconditioner could be applied using N independent parallel processes. We note that this method does not require τ_k to be constant at all time steps and thus could be applied to an adaptive scheme, however the size of the time step at each step would need to be known a priori.

Since each iteration of GMRES or BiCGSTAB is dominated by the work required to complete one solve of a linear system with $\mathscr{P}_{MG}$, the total work at each iteration will be approximately equal to 1N V-cycles. However, as the preconditioner is inherently parallelizable, each time step can be computed on a separate processor. Thus, if distributed over N processors, the total work at each iteration is equivalent to only 1 V-cycle. In order to estimate the convergence of the GMRES iteration we will examine the eigenvalues of the preconditioned system with the exact preconditioner, $\mathscr{P}_{exact}$, defined as

$$\mathscr{P}_{exact} := \text{blkdiag}(M + \tau_1 \theta K, M + \tau_2 \theta K, \ldots, M + \tau_N \theta K) \tag{5}$$

Proposition 1 *Let $\mathscr{A}$ be any block lower triangular matrix and $\mathscr{P}$ is the block diagonal part of $\mathscr{A}$. Then the preconditioned system, $\mathscr{T} = \mathscr{P}^{-1}\mathscr{A}$ where $\mathscr{P}$ is inverted exactly, has eigenvalues all equal to 1 and furthermore, the minimal polynomial is given by*

$$p(\mathscr{T}) = (\mathscr{T} - I)^m \tag{6}$$

for some $m < N$.

Proof We note that $\mathscr{T}$ will be block lower triangular with identity matrices on the diagonal which implies that all the eigenvalues of $\mathscr{T}$ must be equal to 1. Furthermore, the matrix $\mathscr{T} - I$ will be strictly block lower triangular. For any $m < N$, we note that $(\mathscr{T} - I)^m$ will be 0 except for non-zero entries on or below the m-th subdiagonal. Therefore, we must have $(\mathscr{T} - I)^N = 0$ as there are only $N - 1$ subdiagonals.

If we were to calculate the preconditioner exactly, GMRES would converge to the exact solution in at most N steps in exact arithmetic. Numerically we can see that this is still very close to being the case for the multigrid preconditioner defined in (4). Figure 1 shows the convergence for two small system using a Backward Euler discretisation. We can see that there is a sharp drop in the residual when the number of iterations is equal to N for both GMRES and BiCGSTAB iterations. We note that from a linear algebra perspective, this is very interesting as we see that a ‘nearby’ Jordan structure is determining the convergence of a matrix which no longer has the same minimal polynomial. Furthermore, we note that Proposition 1 also applies to any k-step methods and we will provide numerical results for the BDF2 method to support this.

Additionally we see that the multigrid preconditioner, $\mathscr{P}_{MG}$, is spectrally very close to $\mathscr{P}_{exact}$. Figure 2 shows the eigenvalues of the preconditioned system for the Backward Euler ($\theta = 1$) and Crank-Nicolson ($\theta = \frac{1}{2}$) schemes as well as the 2-step BDF method and we see the eigenvalues are closely clustered around 1. Similarly clustered eigenvalues are expected for different parameter values.

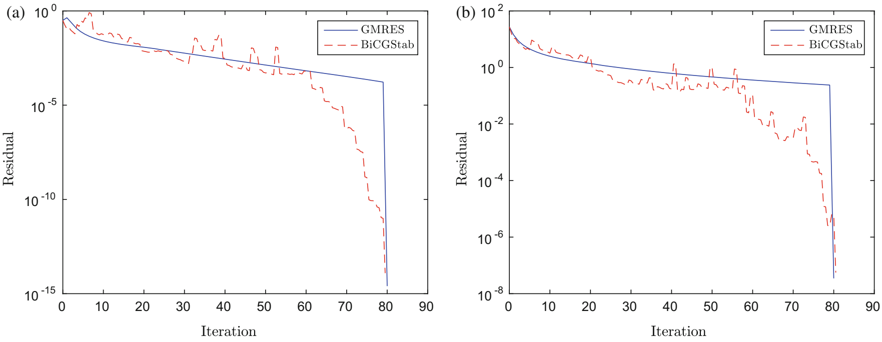


Fig. 1 Convergence of the preconditioned iterative methods ($h = \tau = 2^{-4}, N = 80$). (**a**) Heat equation. (**b**) Convection diffusion

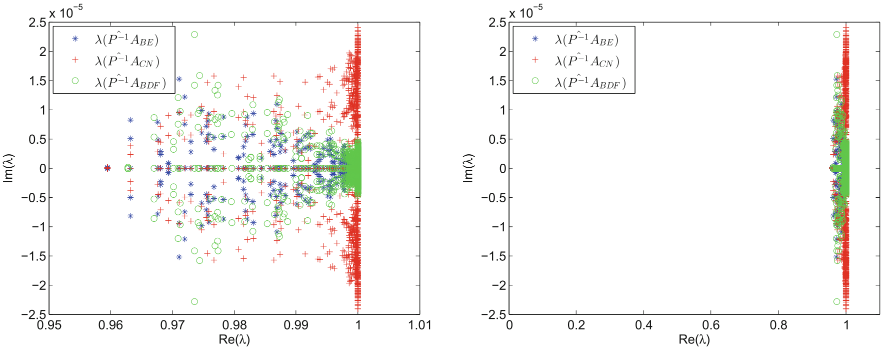


Fig. 2 Eigenvalues of the preconditioned system for the heat equation ($h = \tau = 2^{-5}, N = 5$)

3 Convection-Diffusion Equation

The solutions of the heat equation become exponentially smoother through time so in order to demonstrate that the effectiveness of our proposed method does not rely on this smoothness, we additionally consider the convection-diffusion equation.

The initial-boundary value problem considered is,

$$\begin{aligned} u_t - \epsilon \Delta u + \mathbf{w} \cdot \nabla u &= f && \text{in } \Omega \times (0, T], \quad \Omega \subset \mathbb{R}^2 \text{ or } \mathbb{R}^3, \\ u &= g && \text{on } \partial\Omega, \\ u(x, 0) &= u_0(x) && \text{at } t = 0. \end{aligned} \tag{7}$$

where ϵ is small and positive. As discussed in Chapter 6 of [3], it is often the case that convection effects are more dominant than diffusion. However, in this regime for a Galerkin method, local oscillations can arise in the numerical solution and a stabilisation method is required. We will use the widely employed *Streamline*

Upwind Petrov-Galerkin (SUPG) stabilization method which was introduced in [9] and described for example in [3, Section 6.3.2].

We will use a finite element discretisation in space on a uniform square grid with mesh size h and constant time step τ. Using a general θ-scheme the discretisation and SUPG stabilisation of (7) gives,

$$(M + \tau\theta\hat{K})\mathbf{u}_k = (M - \tau(1-\theta)\hat{K})\mathbf{u}_{k-1} + \tau\mathbf{f}_k \tag{8}$$

where $\hat{K} = \epsilon K + N + \delta S$. Here K represents the stiffness matrix, N represents the discretisation of the convection term $\mathbf{w} \cdot \nabla u$, and S is the stabilisation term. The stabilisation parameter δ required as part of the SUPG method is taken to be the optimal value of 0 when $\mathrm{Pe} \leq 1$ and $\frac{h}{2\|\mathbf{w}\|_2}\left(1 - \frac{1}{\mathrm{Pe}}\right)$ when $\mathrm{Pe} > 1$ where $\mathrm{Pe} = \frac{h\|\mathbf{w}\|_2}{2\epsilon}$.

Due to the formation of layers in the convection-diffusion problem, more care is required in order to use a multigrid approximation. The method we will use is a modified geometric multigrid method described by Ramage in [12] and is specifically designed for convection-diffusion problems.

We now have the following system, $\widehat{\mathscr{A}}\mathbf{u} = \mathbf{f}_{CD}$, where

$$\widehat{\mathscr{A}} = \begin{bmatrix} M + \tau\theta\hat{K} & & & \\ -M + \tau(1-\theta)\hat{K} & M + \tau\theta\hat{K} & & \\ & \ddots & \ddots & \\ & & -M + \tau(1-\theta)\hat{K} & M + \tau\theta\hat{K} \end{bmatrix}, \tag{9}$$

where $\mathbf{u} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_N]^T$ and $\mathbf{f}_{CD} = [M - \tau(1-\theta)\hat{K} + \tau_1\mathbf{f}_1, \tau_2\mathbf{f}_2, \ldots, \tau_N\mathbf{f}_N]^T$ and the $\mathbf{f}_i$ terms contain the extra stabilisation term. We therefore define the approximate preconditioner to be $\widehat{\mathscr{P}}_{MG}^{-1} := \mathrm{blkdiag}((M + \tau\theta\hat{K})_{MG})$, where $(M + \tau\theta\hat{K})_{MG}$ denotes the application of a single Ramage multigrid V-cycle to the matrix $M + \tau\theta\hat{K} \in \mathbb{R}^{n\times n}$. As Proposition 1 still applies to the convection-diffusion system, we expect similar convergence results as were seen in the heat equation.

4 Numerical Results

The results presented in this section were implemented within the IFISS [2] framework. Our implementation of GMRES was from the IFISS package and did not allow restarting. As GMRES can require prohibitive amounts of storage (and growing work) if many iterations are required, we also completed the calculations with the inbuilt Matlab implementation of BiCGSTAB. Both methods were stopped with an absolute residual tolerance of 10^{-6}. The finite element discretisation uses $Q1$ finite elements over the domain.

Heat Equation: For the heat equation, we used the Harwell Subroutine Library AMG preconditioner implementation, HSL_MI20 [1] employed as a "black box".

Table 1 **Heat Equation:** Number of iterations for given grid size and number of time steps

(a) **Smooth test problem**: Backwards Euler

h	τ	**N**	DoF	GMRES	BiCGSTAB
2^{-3}	2^{-3}	40	3240	40	37
2^{-4}	2^{-4}	80	23,120	80	77
2^{-5}	2^{-5}	160	174,240	160	161
2^{-3}	2^{-3}	40	3240	40	37
2^{-4}	2^{-4}	40	11,560	40	39
2^{-5}	2^{-5}	40	43,560	43	43
2^{-6}	2^{-6}	40	169,000	45	45
2^{-7}	2^{-7}	40	665,640	47	45

(b) **Non-smooth test problem:** Crank-Nicholson and BDF2

				Crank-Nicholson		BDF2	
h	τ	**N**	DoF	GMRES	BiCGSTAB	GMRES	BiCGSTAB
2^{-3}	2^{-5}	40	3240	49	48	45	48
2^{-4}	2^{-6}	80	23,120	90	95	87	90
2^{-5}	2^{-7}	160	174,240	175	199	169	190
2^{-3}	2^{-5}	40	3240	49	48	45	48
2^{-4}	2^{-6}	40	11,560	52	50	49	50
2^{-5}	2^{-7}	40	43,560	53	53	51	52
2^{-6}	2^{-8}	40	169,000	56	53	53	55
2^{-7}	2^{-9}	40	665,640	57	54	55	58

We consider two test problems; one with smooth initial data and a second with random initial conditions.

Our first example is defined by the initial conditions, $u_0 = x(x-1)y(y-1)$ with no external forcing (i.e. $f = 0$). Backward Euler was used for the temporal discretisation and the iteration results are summarized in Table 1a below. The second example was defined with random initial data taking values from a uniform distribution on [0, 10]. Crank-Nicholson and BDF2 were used to discretize in time and the results are summarized in Table 1b below. These methods have not been chosen specifically for each problem, but rather just to demonstrate that a variety of time-stepping methods can be used. For both examples it is evident that the iteration numbers are approximately equal to the number of time steps and independent of h.

Convection-Diffusion Equation: For the convection-diffusion equation, the Ramage modified geometric multigrid [12] implemented in IFISS was used. Two pre- and two post-smoothing steps were used with four-directional line Gauss-Seidel smoothing.

In each of the problems, ϵ was set equal to 1/200 so the maximum mesh Peclet number ranged between approximately 46 for $h = 2^{-3}$ and 3 for $h = 2^{-7}$. The convection diffusion test problem is given by Example 6.1.4 in [3] and is known as the double glazing problem. The wind is described by $\mathbf{w} = (2y(1-x^2), -2x(1-y^2))$. Dirichlet boundary conditions are imposed everywhere on the boundary with $\mathbf{u} = 1$

Table 2 Convection-diffusion, double glazing problem. Number of iterations for given grid size and number of steps

h	τ	**N**	DoF	GMRES	BiCGSTAB
2^{-3}	2^{-3}	40	3240	40	41
2^{-4}	2^{-4}	80	23,120	80	78
2^{-5}	2^{-5}	160	174,240	160	167
2^{-4}	2^{-4}	160	46,240	160	182
2^{-5}	2^{-5}	160	174,240	160	167
2^{-6}	2^{-6}	160	676,000	175	176
2^{-7}	2^{-7}	160	2,662,560	165	174

on the boundary where $x = 1$ and zero on all other boundaries. The vector $\mathbf{u}_0$ was zero everywhere except the boundaries where is satisfies the boundary conditions. In Table 2 the iteration numbers remain independent of the mesh size.

5 Conclusions

In this paper we have presented a simple approach for solving the heat equation and the convection-diffusion problem which is parallelizable over time. The method constructs an all-at-once system which is solved using a preconditioned iterative method. The preconditioner is block diagonal and therefore its application can be computed at each time step in parallel over the time steps. We have shown that with the exact preconditioner, GMRES must converge in at most N iterations with exact arithmetic and in practice, the multigrid approximation performs very close to this. Thus, our approach has shown that 'nearby' Jordan structure can be a factor in the convergence of the linear system. Since at each iteration a single AMG V-cycle at each time step can be spread over N processors, N iterations can be completed in approximately the same amount of time as N V-cycles rather than rN V-cycles in a classical sequential approach.

References

1. J. Boyle, M.D. Mihajlovic, J.A. Scott, HSL MI20: an efficient AMG preconditioner. Technical report RAL-TR-2007-021, STFC Rutherford Appleton Laboratory, Didcot (2007)
2. H. Elman, A. Ramage, A. Silvester, Algorithm 866: IFISS, a Matlab toolbox for modelling incompressible flow. ACM Trans. Math. Softw. **33**, 2–14 (2007)
3. H. Elman, D.J. Silvester, A.J. Wathen, *Finite Elements and Fast Iterative Solvers: With Applications in Incompressible Fluid Dynamics*. Numerical Mathematics and Scientific Computation (Oxford University Press, Oxford, 2014)
4. M. Emmett, M.L. Minion, Toward an efficient parallel in time method for partial differential equations. Commun. Appl. Math. Comput. Sci. **7**, 105–132 (2012)

5. M. Gander, Overlapping schwarz for linear and nonlinear parabolic problems, in *9th International Conference on Domain Decomposition*, Bergen, 1996, ed. by P.E. Bjørstad, M.S. Espedal, D.E. Keyes
6. M.J. Gander, 50 years of time parallel time integration. Technical report, University of Geneva (2014)
7. M.J. Gander, S. Güttel, PARAEXP: a parallel integrator for linear initial-value problems. SIAM J. Sci. Comput. **35**, 123–142 (2013)
8. G. Horton, S. Vandewalle, A space-time multigrid method for parabolic partial differential equations. SIAM J. Sci. Comput. **16**, 848–864 (1995)
9. T.J.R. Hughes, A. Brooks, A multidimensional upwind scheme with no crosswind diffusion. Finite Element Methods Convect. Domin. Flows **34**, 19–35 (1979)
10. J. Lions, Y. Maday, G. Turinici, A 'parareal' in time discretization of PDE's. Comptes Rendus de l'Acad. des Sci. Ser. I Math. **332**, 661–668 (2001)
11. M. Neumüller, Space-time methods: fast solvers and applications. Ph.D. thesis, Graz University of Technology (2013)
12. A. Ramage, A multigrid preconditioner for stabilised discretisation of advection-diffusion problems. J. Comput. Appl. Math. **110**, 187–203 (1999)
13. Y. Saad, M.H. Schultz, GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems. SIAM J. Sci. Stat. Comput. **7**(3), 856–869 (1986)
14. H. Van der Vorst, Bi-CGSTAB: a fast and smoothly convergent variant of Bi-CG for the solution of nonsymmetric linear systems. SIAM J. Sci. Stat. Comput. **13**(2), 631–644 (1992)

The Induced Dimension Reduction Method Applied to Convection-Diffusion-Reaction Problems

Reinaldo Astudillo and Martin B. van Gijzen

Abstract Discretization of (linearized) convection-diffusion-reaction problems yields a large and sparse non symmetric linear system of equations,

$$A\mathbf{x} = \mathbf{b}. \tag{1}$$

In this work, we compare the computational behavior of the Induced Dimension Reduction method (IDR(s)) (Sonneveld and van Gijzen, SIAM J Sci Comput 31(2):1035–1062, 2008), with other short-recurrences Krylov methods, specifically the Bi-Conjugate Gradient Method (Bi-CG) (Fletcher, Conjugate gradient methods for indefinite systems. In: Proceedings of the Dundee conference on numerical analysis, pp 73–89, 1976), restarted Generalized Minimal Residual (GMRES(m)) (Saad and Schultz, SIAM J Sci Stat Comput 7:856–869, 1986), and Bi-Conjugate Gradient Stabilized method (Bi-CGSTAB) (van der Vorst, SIAM J Sci Stat Comput 13(2):631–644, 1992).

1 Introduction

In this paper we consider the following simple convection-diffusion-reaction model problem

$$-\epsilon \triangle u + \mathbf{v}^T \nabla u + \rho u = f, \qquad \text{in } \Omega = [0,\ 1]^d \tag{2}$$

with $d = 2$ or $d = 3$, and Dirichlet boundary conditions $u = 0$ on $\partial\Omega$. In Eq. (2), u represents the concentration of solute, $\mathbf{v} \in \mathbb{R}^d$ is the velocity of the medium or convection vector, $\epsilon > 0$ represents the diffusion coefficient, ρ the reaction coefficient, and f represents the source-term function.

R. Astudillo (✉) • M.B. van Gijzen
Delft Institute of Applied Mathematics, Delft University of Technology, Mekelweg 4, 2629 CD, Delft, The Netherlands
e-mail: r.a.astudillo@tudelft.nl; m.b.vangijzen@tudelft.nl

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_29

Discretization of the Eq. (2) yields a non-symmetric system of linear equations,

$$A\mathbf{x} = \mathbf{b}, \tag{3}$$

where $\mathbf{x}$ is the unknown vector in $\mathbb{R}^N$, $\mathbf{b} \in \mathbb{R}^N$, and $A \in \mathbb{R}^{N\times N}$ is typically large, and sparse. Krylov subspace methods are a popular choice to solve such systems. However, the convergence ratio of these methods are strongly influenced by the numerical properties of the coefficient matrix A, which internally depend on the physical parameters of Eq. (2). For example, in the convection-dominated case, i.e. $\|\mathbf{v}\| >> \epsilon$, the coefficient matrix A has almost purely imaginary eigenvalues and this can slow down the convergence of Krylov methods.

GMRES [3] is an optimal method, it obtains the best approximation in a subspace of dimension j performing j matrix-vector products. Nevertheless, due the large and ill-conditioned linear systems obtained from the discretization of the convection-diffusion-reaction equation, one can expect that many iterations need to be performed to compute the solution accurately. For this reason and taking into account that the computational cost of GMRES increases per iteration, it is preferable to use a preconditioned short-recurrences Krylov method that keeps the computational work and memory consumption fixed per iteration. Bi-CGSTAB [10] is the most widely used method of this kind. However, IDR(s) outperforms Bi-CGSTAB in the experiments presented in [9] and [11]. In this work we continue this comparison. We compare the numerical behavior of Bi-CG [1], GMRES(m), Bi-CGSTAB, and IDR(s) to solve the linear systems arising from the discretization of (2).

2 Krylov Methods for Solving Systems of Linear Equations

A projection method onto m-dimensional subspace $\hat{\mathscr{K}}$ and orthogonal to the m-dimensional subspace $\mathscr{L}$, is an iterative method to solve (3) which finds the approximate solution $\mathbf{x}_m$ in the affine subspace $\mathbf{x}_0 + \hat{\mathscr{K}}$ imposing the Petrov-Galerkin condition, i.e., $\mathbf{r}_m = \mathbf{b} - A\mathbf{x}_m$ orthogonal to $\mathscr{L}$. The subspace $\hat{\mathscr{K}}$ is called search space, while $\mathscr{L}$ is called restriction space.

The Krylov subspace methods are projection methods for which the search space is the Krylov subspace $\mathscr{K}_m(A, \mathbf{r}_0) = \text{span}\{\mathbf{r}_0, A\mathbf{r}_0, \ldots, A^{m-1}\mathbf{r}_0\}$, where $\mathbf{r}_0 = \mathbf{b} - A\mathbf{x}_0$ with $\mathbf{x}_0$ as initial guess in $\mathbb{C}^N$. The different Krylov methods are obtained from the different choices of the restriction space. For a comprehensive description of the Bi-CG, GMRES(m) and Bi-CGSTAB, we refer the reader to [2].

2.1 The Induction Dimension Reduction Method (IDR(s))

IDR(s) was introduced in 2008 [9]. This method can also be described as a Krylov projection method (see [6]), however, the original formulation and implementation of IDR(s) is based on the following theorem.

Theorem 1 (IDR theorem) *Let A be any matrix in $\mathbb{C}^{N\times N}$, and let $P = [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_s]$ be an $N \times s$ matrix with s linear independent columns. Let $\{\mu_j\}$ be a sequence in $\mathbb{C}$. With $\mathscr{G}_0 \equiv \mathbb{C}^N$, define*

$$\mathscr{G}_{j+1} \equiv (A - \mu_{j+1}I)(\mathscr{G}_j \cap P^{\perp}) \quad j = 0, 1, 2 \ldots,$$

where $P^{\perp}$ represents the orthogonal complement of P. If $P^{\perp}$ does not contain an eigenvector of A, then, for all $j = 0, 1, 2 \ldots$, the following hold

1. *$\mathscr{G}_{j+1} \subset \mathscr{G}_j$, and*
2. *dimension($\mathscr{G}_{j+1}$) $<$ dimension($\mathscr{G}_j$) unless $\mathscr{G}_j = \{\mathbf{0}\}$.*

Proof See [9].

Assuming that $s+1$ approximations are available with their corresponding residuals belonging to $\mathscr{G}_{j-1}$, IDR(s) constructs the new approximation $\mathbf{x}_k$ at the iteration k, imposing the condition that the vector $\mathbf{r}_k = \mathbf{b} - A\mathbf{x}_k$ should be in the subspace $\mathscr{G}_j$. Moreover, using the fact that $\mathscr{G}_j \subset \mathscr{G}_{j-1}$, IDR($s$) creates inductively $s+1$ residuals in the subspace $\mathscr{G}_j$. After this, it is possible to create new residuals in $\mathscr{G}_{j+1}$.

IDR(s) has three attractive numerical properties. First, IDR(s) uses recurrences of size $s+1$, and the parameter s is normally selected between 2 and 8. Second, the subspaces $\mathscr{G}_j$ with $j = 1, 2, \ldots$ are nested and shrinking, and for this reason, IDR(s) has guarantee convergence in at most $N + \frac{N}{s}$ matrix-vector multiplication in exact arithmetic (see Corollary 3.2 in [9]). Third, IDR(1) and Bi-CGSTAB are mathematically equivalent (see [5]); and IDR(s) is commonly faster than Bi-CGSTAB for $s > 1$. The details of the implementation of IDR(s) can be found in [11].

3 Numerical Experiments

All the experiment presented in this section are the discretization of Eq. (2) with homogeneous Dirichlet boundary conditions over the unit cube, The right-hand-side function f is defined by the solution $u(x, y, z) = x(1-x)y(1-y)z(1-z)$. We use as stopping criteria that,

$$\frac{\|\mathbf{b} - A\mathbf{x}_k\|_2}{\|\mathbf{b}\|_2} < 10^{-8}.$$

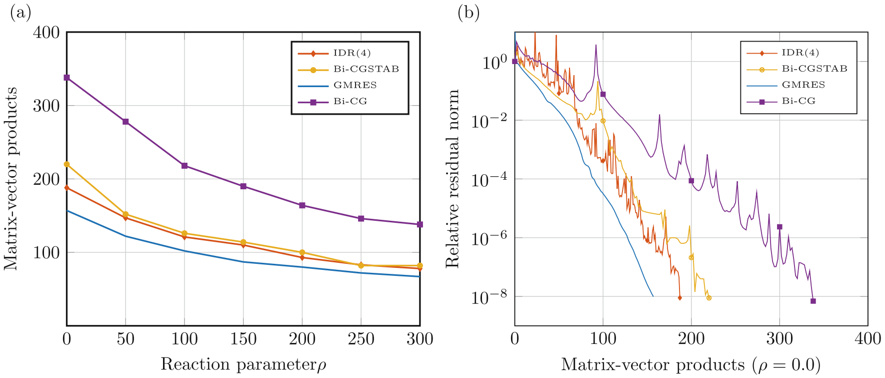


Fig. 1 *Example 1:* (**a**) Number of matrix-vector products required to converge as a function of the parameter ρ for a diffusion-dominated problem. (**b**) Comparison of the residual norms, the physical parameters are $\epsilon = 1.0$, $\mathbf{v} = (1.0,\ 1.0,\ 1.0)^T/\sqrt{3}$, and $\rho = 0.0$

The discretization of Eq. (2) using central finite differences may produce unphysical oscillations in the numerical solution of convection or reaction dominated problems. This problem can be solved discretizing the convection term using upwind schemes. However, we use central finite differences rather than upwind dicretization in this set of problems, in order to illustrate the effect of unfavorable numerical conditions over the Krylov subspace solvers.

Experiment 1 In this example, we consider the parameters $\epsilon = 1.0$ and $\mathbf{v} = (1.0,\ 1.0,\ 1.0)^T/\sqrt{3}$. We want to illustrate the effect of non-negative reaction parameter over the Krylov solver, then, we select $\rho \in \{0, 50, \ldots, 300\}$. Figure 1a shows the number of matrix-vector multiplication required for each Krylov method as a function of the reaction parameter ρ. In these problems, the increment of the reaction parameter produces a reduction in the number of matrix-vector products required for each Krylov method. All the methods perform very efficiently for these examples. Figure 1b shows the evolution of the residual norm for $\rho = 0.0$. The execution times are: IDR(4) 0.62s, Bi-CGSTAB 0.64s, Bi-CG 0.92s, and GMRES 2.83s.

Experiment 2 In order to illustrate the effect of the magnitude of the convection velocity, we consider $\epsilon = 1.0$, $\rho = -50.0$, and $\mathbf{v} = \beta(1.0,\ 1.0,\ 1.0)^T/\sqrt{3}$ with $\beta \in \{100.0,\ 200.0,\ \ldots, 800.0\}$. As the parameter β grows we obtain a more convection-dominated problem. Figure 2a shows how many matrix-vector products are required for each Krylov method as function of the convection speed. The problem is more convection-dominated as $\|\mathbf{v}\|_2$ grows. It is interesting to remark the linear of the number of matrix-vector product for Bi-CGSTAB. Figure 1b shows the evolution of the residual norm for $\beta = 800.0$. Execution time IDR(4) 1.24s, Bi-CGSTAB 5.64s, Bi-CG 1.01s, and GMRES 3.26s.

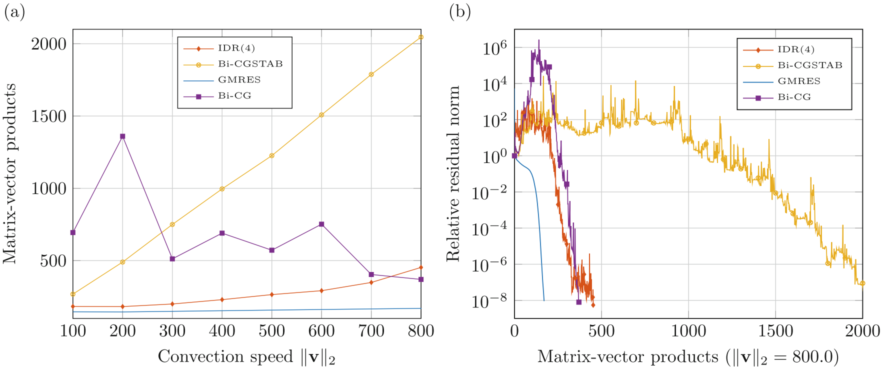


Fig. 2 *Example 2:* (**a**) Number of matrix-vector products required to converge as a function of the convection speed. (**b**) Comparison of the residual norms, the physical parameters are $\epsilon = 1.0$, $\mathbf{v} = 800.0 \times (1.0,\ 1.0,\ 1.0)^T/\sqrt{3}$, and $\rho = -50.0$

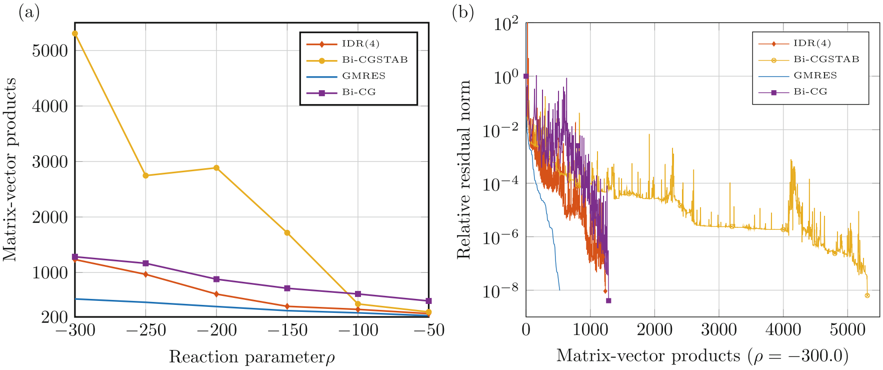


Fig. 3 *Example 3:* (**a**) Number of matrix-vector products required to converge as a function of the parameter ρ. (**b**) Comparison of the residual norms. The physical parameters $\epsilon = 1$, $\mathbf{v} = (1.0,\ 1.0,\ 1.0)^T/\sqrt{3}$, and $\rho = -300.0$

Experiment 3 Here we use the same set of problems presented in experiment 1, but selecting negative reaction parameters, we consider $\rho \in \{-300, 250, \ldots, -50\}$. In Figure 3a, one can see how the negative of the reaction parameter generates a considerable increment of the matrix-vector needed for solving the corresponding linear system. Bi-CGSTAB perform poorly for large negative reaction parameter. Figure 1b shows the evolution of the residual norm for $\epsilon = 1$ and $\rho = 300.0$. The execution time are: IDR(4) 4.02s, Bi-CGSTAB 15.38s, Bi-CG 3.52 s, and GMRES 28.57s.

3.1 IDR(s) and Bi-CG

Despite being a method that is not drastically affected by the increment of the reaction parameter or the convection speed, Bi-CG is not the faster method in terms matrix-vector products required. Bi-CG requires two matrix-vector multiplications to produce one new approximation. IDR(4) in most of the experiments requires less matrix-vector multiplication to get the desired residual tolerance. Only in the highly convection-dominated examples presented in the experiment 2, Bi-CG presents a similar behavior as IDR(4). A discussion of the phenomena is presented in Sect. 3.3.

3.2 IDR(s), GMRES, and Restarted GMRES

In the numerical experiments presented in the previous section, Full GMRES is the methods that uses less matrix-vector products to obtain the desired residual reduction. This result is expected due the optimal residual condition of GMRES. Nevertheless, the computational requirements of full GMRES grow in every iteration. Restarting GMRES or GMRES(m) is an option to overcome this issue. The idea of GMRES(m) is to limit to a maximum of m matrix-vector products, and then restart the process using the last approximation as initial vector. The optimal residual property is lost in this restarting scheme.

In terms of memory consumption, GMRES(m) is equivalent to IDR(s) when $m = 3(s+1)$. In order to compare the behavior of GMRES(m) and IDR(4), we consider the discretization of Eq. (2) with this parameters: $\epsilon = 1$, $\mathbf{v} = (1.0,\ 1.0,\ 1.0)^T/\sqrt{3}$ and $\rho = 40.0$, and we take as restarting parameter $m = 15, 16, \ldots, 170$. Figure 4 shows the number of matrix-vector multiplication required for GMRES(m) for different values of m. GMRES(160) and IDR(4) solve this system using the same number of matrix-vector products (262), however, GMRES(160) consumes approximately ten times more memory than IDR(4). Moreover, CPU time for GMRES(160) is 4.60s while IDR(4) runs in only 0.79s.

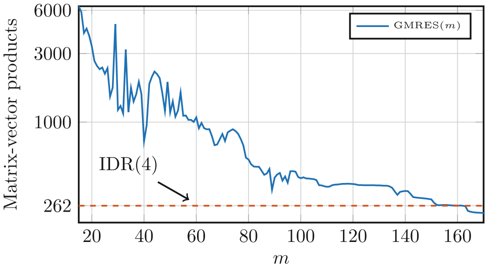


Fig. 4 (GMRES(m) and IDR(s) comparison) Number of matrix-vector products required for GMRES(m) as a function of the parameter m

3.3 IDR(s) and Bi-CGSTAB

One can see in the experiments that Bi-CGSTAB performs poorly for convection-dominated problems. This can be explained throughout the study of the residual formulas for Bi-CGSTAB. The residual vector in Bi-CGSTAB can be written in the form,

$$\mathbf{r}_k^{(B)} = \Omega_k(A)\phi_k(A)\mathbf{r}_0,$$

where $\phi_k(t)$ is residual associated with Bi-CG and $\Omega_k(t)$ is the Minimal Residual (MR) polynomial defined as,

$$\Omega_k(t) = (1 - \omega_k t)\Omega_{k-1}(t).$$

The parameter ω_k are selected such that $\|\mathbf{r}_k^{(B)}\|_2$ is minimized. However, for indefinite matrices or real matrices that have non-real eigenvalues with an imaginary part that is large relative to the real part, the parameter ω_k is close to zero (see [7]), and the MR-polynomial suffers from slow convergence or numerical instability. To illustrate this we show the behavior of the polynomial $\Omega_k(A)$ applied to two different matrices from the second set of experiments. We consider $\beta = 100.0$ and $\beta = 800.0$ labeled in Fig. 5 as moderate convection-dominated and highly convection-dominated respectively.

IDR(s) and Bi-CGSTAB are closely related, in fact, Bi-CGSTAB and IDR(1) are mathematically equivalent for the same parameter choice (see [5]). The convergence of IDR is also affected by the convection speed for a similar reason. The IDR(s)

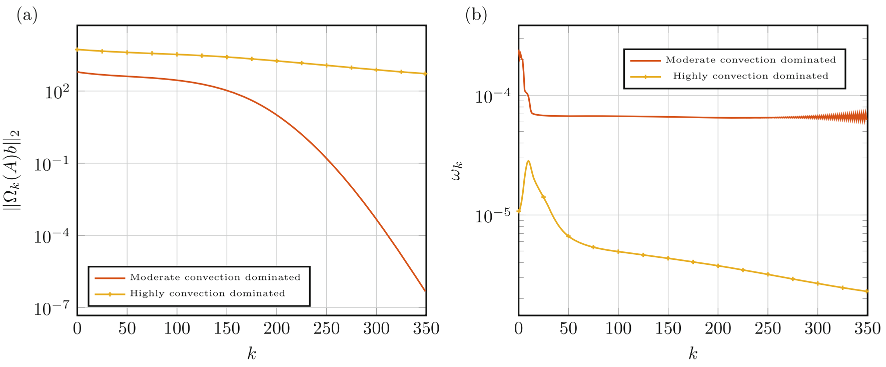


Fig. 5 (**a**) Behavior of the norm of the MR-polynomial $\Omega_k(A)$. (**b**) Values of the parameter ω_k

residual vector $\mathbf{r}_k$ in the subspace $\mathscr{G}_j$ can be written as,

$$\mathbf{r}_k^{(I)} = \Omega_j(A)\psi_{k-j}(A)\mathbf{r}_0,$$

where $\psi_{k-j}(t)$ is a block Lanczos polynomial. For IDR(s) the degree of the polynomial $\Omega_k(t)$ increases by one every $s+1$ matrix-vector products, while in Bi-CGSTAB this degree grows by one every two matrix vector products. For this reason, IDR(s) controls the negative effects of the MR-polynomial when A has complex spectrum or is an indefinite matrix.

The bad convergence for strongly convection-dominated problems of Bi-CGSTAB has been observed by several authors and has given rise to BiCGstab(ℓ) [4]. This method uses polynomial factors of degree ℓ, instead of MR-polynomial. A similar strategy has been implemented in IDR(s) which led to the method IDRstab [8]. For the comparison of the convergence of BiCGstab(ℓ) and IDRstab with IDR(s) we refer the reader to [8].

4 Conclusions

Throughout the numerical experiment, we have shown that IDR(s) is a competitive option to solve system of linear equation arising in the discretization of the convection-diffusion-reaction equation.

GMRES, Bi-CG, and IDR(s) exhibit a stable behavior in the most numerically difficult examples conducted in this work. Despite performing more matrix-vector products to obtain convergence, IDR(s) consumes less CPU time than GMRES. We show that for diffusion-dominated problems with a positive reaction term the convergence of the Bi-CGSTAB and IDR(s) are very similar, and for this kind of problems it is often preferable to simply choose $s = 1$. However, for the more difficult to solve convection dominated problems, or problems with a negative reaction term, IDR(s), with $s > 1$ greatly outperform Bi-CGSTAB.

References

1. R. Fletcher, Conjugate gradient methods for indefinite systems, ed. by G.A. Watson. *Proceedings of the Dundee Biennal Conference on Numerical Analysis 1974* (Springer, New York, 1975), pp. 73–89
2. Y. Saad, *Iterative Methods for Sparse Linear Systems*, 2nd edn. (Society for Industrial and Applied Mathematics, Philadelphia, 2003)
3. Y. Saad, M. Schultz, GMRES: a generalized minimal residual algorithm for solving nonsymetric linear systems. SIAM J. Sci. Stat. Comput. **7**, 856–869 (1986)
4. G.L.G. Sleijpen, D.R. Fokkema, BiCGstab(ℓ) for linear equations involving Unsymmetric matrices with complex spectrum. Electron. Trans. Numer. Anal. **1**, 11–32 (1993)

5. G.L.G. Sleijpen, P. Sonneveld, M.B. van Gijzen, Bi-CGSTAB as an induced dimension reduction method. Appl. Numer. Math. **60**, 1100–1114 (2010)
6. V. Simoncini, D.B. Szyld, Interpreting IDR as a Petrov-Galerkin method. SIAM J. Sci. Comput. **32**(4), 1898–1912 (2010)
7. G.L.G. Sleijpen, H.A. van der Vorst, Maintaining convergence properties of bicgstab methods in finite precision arithmetic. Numer. Algorithms **10**(2), 203–223 (1995)
8. G.L.G. Sleijpen, M.B. van Gijzen, Exploiting BiCGstab(ℓ) strategies to induce dimension reduction. SIAM J. Sci. Comput. **32**(5), 2687–2709 (2010)
9. P. Sonneveld, M.B. van Gijzen, IDR(s): a family of simple and fast algorithms for solving large nonsymmetric systems of linear equations. SIAM J. Sci. Comput. **31**(2), 1035–1062 (2008)
10. H.A. van der Vorst, Bi-CGSTAB: a fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems. SIAM J. Sci. Stat. Comput. **13**(2), 631–644 (1992)
11. M.B. van Gijzen, P. Sonneveld, Algorithm 913: an elegant IDR(s) variant that efficiently exploits bi-orthogonality properties. ACM Trans. Math. Softw. **38**(1), 5:1–5:19 (2011)

Block Variants of the COCG and COCR Methods for Solving Complex Symmetric Linear Systems with Multiple Right-Hand Sides

Xian-Ming Gu, Bruno Carpentieri, Ting-Zhu Huang, and Jing Meng

Abstract In the present study, we establish two new block variants of the Conjugate Orthogonal Conjugate Gradient (COCG) and the Conjugate A-Orthogonal Conjugate Residual (COCR) Krylov subspace methods for solving complex symmetric linear systems with multiple right hand sides. The proposed Block iterative solvers can fully exploit the complex symmetry property of coefficient matrix of the linear system. We report on extensive numerical experiments to show the favourable convergence properties of our newly developed Block algorithms for solving realistic electromagnetic simulations.

1 Introduction

In this paper we are interested in the efficient solution of linear systems with multiple right-hand sides (RHSs) of the form

$$AX = B, \qquad A \in \mathbb{C}^{n\times n},\ \ X, B \in \mathbb{C}^{n\times p},\ p \ll n, \tag{1}$$

X.-M. Gu • T.-Z. Huang
School of Mathematical Sciences, University of Electronic Science and Technology of China, Chengdu 611731, P.R. China
e-mail: guxianming@live.cn; tingzhuhuang@126.com

X.-M. Gu
Johann Bernoulli Institute for Mathematics and Computer Science, University of Groningen, Nijenborgh 9, P.O. Box 407, 9700 AK Groningen, The Netherlands

B. Carpentieri (✉)
School of Science and Technology, Nottingham Trent University, Clifton Campus, Nottingham, NG11 8NS, UK
e-mail: bruno.carpentieri@ntu.ac.uk

J. Meng
School of Mathematics and Statistics, Taishan University, Taian 271021, P.R. China
e-mail: mengmeng-erni@163.com

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_30

where A is a non-Hermitian but symmetric matrix, i.e., $A \neq A^H$ and $A = A^T$. Linear systems of this form arise frequently in electromagnetic scattering applications, for example in monostatic radar cross-section calculation, where each right-hand side typically corresponds to an incident wave illuminating the target at a given angle of incidence [1, 2].

Roughly speaking, computational techniques for solving linear systems on modern computers can be divided into the class of direct and of iterative methods. Block iterative Krylov subspace methods are particularly designed for solving efficiently linear systems with multiple RHSs (cf. [3, 4]). Block algorithms require one or more matrix product operations of the form AV, with $V \in \mathbb{C}^{n\times p}$ an arbitrary rectangular matrix, per iteration step. Thus they can solve the typical memory bottlenecks of direct methods. However, most of them, such as the Block Bi-Conjugate Gradient (bl_bicg) [5], Block Bi-Conjugate Residual (bl_bicr) [3], Block BiCGSTAB (bl_bicgstab) [6], Block BiCRSTAB (bl_bicrstab) [3], Block QMR (bl_qmr) [7], Block IDR(s) (bl_idr(s)) [8] and Block GMRES (bl_gmres) [9] methods, do not naturally exploit any symmetry of A.

Methods that can exploit the symmetry of A are typically of (quasi) minimal residual type (i.e. bl_sqmr) [7]. Tadano and Sakurai recently proposed the Block COCG (bl_cocg) [10] method, which can be regarded as a natural extension of the COCG [11] algorithm for solving linear systems (1). Both these two methods need one operation AV per iteration step. In this paper we revisit the Block COCG method, presenting a more systematic derivation than the one presented [10], and we introduce a new Block solver (bl_cocr) that can be seen as an extension of the COCR algorithm proposed in [12]. The numerical stability of the bl_cocg and the bl_cocr methods are enhanced by the residual orthonormalization technique [13].

The paper is organized as follows. In Sect. 2 we present the general framework for the development of the bl_cocg and the bl_cocr solvers. In Sect. 3 we study their numerical stability properties and then we show how to improve their convergence by employing the residual orthonormalization technique. In Sect. 3, we report on extensive numerical experiments to illustrate the effectiveness of the two new iterative methods in computational electromagnetics. Finally, some conclusions arising from this work are presented in Sect. 4.

2 The Block COCG and Block COCR Methods

Let $X^{m+1} \in \mathbb{C}^{n\times p}$ be the $(m+1)$th approximate solution of linear systems (1) satisfying the following condition

$$X_{m+1} = X_0 + Z_{m+1}, \quad Z_{m+1} \in \mathscr{K}^{\diamond}_{m+1}(A; R_0), \tag{2}$$

where $R_0 = B - AX_0$ is an initial residual and $\mathscr{K}^{\diamond}_{m+1}(A; R_0)$ is the block Krylov subspace [4] defined as

$$\mathscr{K}^{\diamond}_{m+1}(A; R_0) = \Big\{ \sum_{j=0}^{m} A^j R_0 \gamma_j \mid \gamma_j \in \mathbb{C}^{p \times p} \ (j = 0, 1, \ldots, m) \Big\}. \tag{3}$$

Compared with conventional Krylov subspace methods, where $\boldsymbol{x}^{(j)}_{m+1} - \boldsymbol{x}^{(j)}_0 \in \mathscr{K}_{m+1}(A, \boldsymbol{r}^{(j)}_0)$, note that block Krylov methods can search the approximate solutions into larger spaces, and thus they may require less iterations to converge to a given accuracy. In the next section we introduce the framework for the development of the Block COCG and the Block COCR methods.

2.1 Derivation of the Block COCG and Block COCR Methods

According to Eqs. (2) and (3), the $(m+1)$th residual $R_{m+1} = B - AX_{m+1}$ of the Block COCG method [10] and the Block COCR method is computed by the following recurrence relations,

$$\begin{aligned} R_0 = P_0 = B - AX_0 &\in \mathscr{K}^{\diamond}_1(A; R_0), \\ R_{m+1} = R_m - AP_m\alpha_m &\in \mathscr{K}^{\diamond}_{m+2}(A; R_0), \\ P_{m+1} = R_{m+1} + P_m\beta_m &\in \mathscr{K}^{\diamond}_{m+2}(A; R_0). \end{aligned} \tag{4}$$

Here, $P_{m+1} \in \mathbb{C}^{n \times p}, \alpha_m, \beta_m \in \mathbb{C}^{p \times p}$. The $(m+1)$th approximate solution X_{m+1} is updated through the recurrence relation

$$X_{m+1} = X_m + P_m\alpha_m. \tag{5}$$

Similarly to the framework introduced in [14], different formulae for the $p \times p$ matrices α_m, β_m $(m = 0, 1, \ldots)$ in the recurrences (4) and (5) lead to different iterative algorithms. Denoting by $\mathscr{L}$ the *block constraints subspace*, these matrices α_m, β_m are determined by imposing the orthogonality conditions

$$R_m \perp \mathscr{L} \quad \text{and} \quad AP_m \perp \mathscr{L}. \tag{6}$$

The Block COCG and the Block COCR methods correspond to the choices $\mathscr{L} = \mathscr{K}^{\diamond}_m(\bar{A}; \bar{R}_0)$ and $\mathscr{L} = \bar{A}\mathscr{K}^{\diamond}_m(\bar{A}; \bar{R}_0)$, respectively. In Table 1, the conjugate orthogonality conditions imposed to determine α_m and β_m are summarized for the sake of clarity.

We show the complete Block COCR algorithm in Algorithm 1. We use the notation $\|\cdot\|_F$ for the Frobenius norm of a matrix, and ϵ is a sufficiently small user-defined value. We see that the Block COCR method requires two matrix products

Table 1 Orthogonality conditions imposed to determine $p \times p$ matrices α_m, β_m

Matrix	Block COCG	Block COCR
α_m, β_m	$R_m \perp \mathscr{K}_m^{\diamond}(\bar{A}; \bar{R}_0)$	$R_m \perp \bar{A}\mathscr{K}_m^{\diamond}(\bar{A}; \bar{R}_0)$
	$AP_m \perp \mathscr{K}_m^{\diamond}(\bar{A}; \bar{R}_0)$	$AP_m \perp \bar{A}\mathscr{K}_m^{\diamond}(\bar{A}; \bar{R}_0)$

Algorithm 1 The Block COCR method

```
1: $X_0 \in \mathbb{C}^{n \times p}$ is an initial guess, $R_0 = B - AX_0$,
2: Set $P_0 = R_0$, $U_0 = V_0 = AR_0$,
3: for $m = 0, 1, \ldots$, until $\|R_m\|_F / \|R_0\|_F \leq \epsilon$ do
4:     Solve $(U_m^T U_m)\alpha_m = R_m^T V_m$ for $\alpha_m$,
5:     $X_{m+1} = X_m + P_m \alpha_m$,
6:     $R_{m+1} = R_m - U_m \alpha_m$ and $V_{m+1} = AR_{m+1}$,
7:     Solve $(R_m^T V_m)\beta_m = R_{m+1}^T V_{m+1}$ for $\beta_m$,
8:     $P_{m+1} = R_{m+1} + P_m \beta_m$,
9:     $U_{m+1} = V_{m+1} + U_m \beta_m$,
10: end for
```

AP_{m+1}, AR_{m+1} at each iteration step. While the product AR_{m+1} is computed by explicit matrix multiplication, the product AP_{m+1} is computed by the recurrence relation at line 9, to reduce the computational complexity. Note that the Block COCG and the Block COCR methods can be derived from the Block BiCG and the Block BiCR methods, respectively, by choosing the initial auxiliary residual $\hat{R}_0 = \bar{R}_0$ and removing some redundant computations; we refer to the recent work [14] for similar discussions about the derivation of conventional non-block Krylov subspace methods for complex symmetric linear systems with single RHS.

2.2 *Improving the Numerical Stability of the Block COCG and Block COCR Methods by Residual Orthonormalization*

One known problem with Block Krylov subspace methods is that the residual norms may not converge when the number p of right-hand sides is large, mainly due to numerical instabilities, see e.g. [13]. These instabilities often arise because of the loss of linear independence among the column vectors of the $n \times p$ matrices that appear in the methods, such as R_m and P_m. Motivated by this concern, in this section we propose to use the residual orthonormalization technique to enhance the numerical stability of the Block COCG and Block COCR algorithms. This efficient technique was introduced in [13] in the context of the Block CG method [5].

Let the Block residual R_m be factored as $R_m = Q_m \xi_m$ by conventional QR factorization,[1] with $Q_m^H Q_m = I_p$. Here I_p denotes the identity matrix of order p

[1] For our practical implementation, we use MATLAB qr-function "`qr`(W,0)" for a given matrix $W \in \mathbb{C}^{n \times p}$.

Algorithm 2 Algorithm of the Block COCG method with residual orthonormalization (bl_cocg_rq)

```
$X_0 \in \mathbb{C}^{n\times p}$ is an initial guess, and compute $Q_0\xi_0 = B - AX_0$,
Set $S_0 = Q_0$,
for $m = 0, 1, \ldots,$ until $\|\xi_m\|_F/\|B\|_F \leq \epsilon$ do
    Solve $(S_m^T A S_m)\alpha'_m = Q_m^T Q_m$ for $\alpha'_m$,
    $X_{m+1} = X_m + S_m\alpha'_m\xi_m$,
    $Q_{m+1}\tau_{m+1} = Q_m - AS_m\alpha'_m$ and $\xi_{m+1} = \tau_{m+1}\xi_m$,
    Solve $(Q_m^T Q_m)\beta'_m = \tau_{m+1}^T Q_{m+1}^T Q_{m+1}$ for $\beta'_m$,
    $S_{m+1} = Q_{m+1} + S_m\beta'_m$,
end for
```

Algorithm 3 Algorithm of the Block COCR method with residual orthonormalization (bl_cocr_rq)

```
$X_0 \in \mathbb{C}^{n\times p}$ is an initial guess, and compute $Q_0\xi_0 = B - AX_0$,
Set $S_0 = Q_0$ and $U_0 = V_0 = AQ_0$,
for $m = 0, 1, \ldots,$ until $\|\xi_m\|_F/\|B\|_F \leq \epsilon$ do
    Solve $(U_m^T U_m)\alpha'_m = Q_m^T U_m$ for $\alpha'_m$,
    $X_{m+1} = X_m + P_m\alpha'_m$
    $Q_{m+1}\tau_{m+1} = Q_m - U_m\alpha'_m$ and $\xi_{m+1} = \tau_{m+1}\xi_m$,
    Compute $V_{m+1} = AQ_{m+1}$,
    Solve $(Q_m^T V_m)\beta_m = \tau_{m+1}^T Q_{m+1}^T V_{m+1}$ for $\beta'_m$,
    $S_{m+1} = Q_{m+1} + S_m\beta'_m$,
    $U_{m+1} = V_{m+1} + U_m\beta'_m$,
end for
```

and $\xi_m \in \mathbb{C}^{p\times p}$. From (4), the following equation can be obtained

$$Q_{m+1}\tau_{m+1} = Q_m - AS_m\alpha'_k. \tag{7}$$

Here, $\tau_{m+1} \equiv \xi_{m+1}\xi_{m-1}$, $\alpha'_k \equiv \xi_m\alpha_m\xi_{m-1}$, and $S_m = P_m\xi_{m-1}$. In the new Algorithms 2 and 3, the matrix β'_m is defined as $\alpha'_m \equiv \xi_m\beta_m\xi_{m+1}^{-1}$. The residual norm is monitored by $\|\xi_m\|_F$ instead of $\|R_m\|_F$, since the Frobenius norm of R_m satisfies $\|R_m\|_F = \|\xi_m\|_F$. Note that the QR decomposition is calculated at each iteration. However, the numerical results shown in the next section indicate that the extra cost is amortized by the improved robustness of the two Block solvers.

3 Numerical Experiments

In this section, we carry out some numerical experiments to show the potential effectiveness of the proposed iterative solution strategies in computational electromagnetics. We compare the bl_cocg, bl_cocg_rq, bl_cocr, bl_cocr_rq methods against other popular block Krylov subspace methods such as bl_qmr, bl_bicgstab, bl_bicrstab, bl_idr(s) (selecting matrix $P = rand(n, sp)$, see [8]) and restarted

Table 2 The numerical results of different iterative solvers for the first example. The solution time refers to the wall clock time expressed in seconds

	young2c ($p = 10$)			young3c ($p = 8$)			young1c ($p = 8$)		
Method	*Iters*	*TRR*	Time	*Iters*	*TRR*	Time	*Iters*	*TRR*	Time
bl_cocg	238	−10.03	0.17	†	†	†	329	−10.16	0.16
bl_cocg_rq	142	−10.14	0.13	151	−10.00	0.09	177	−10.29	0.12
bl_cocr	201	−10.07	0.15	145	−9.95	0.04	221	−10.07	0.12
bl_cocr_rq	138	−10.18	0.13	146	−10.03	0.05	180	−10.18	0.13
bl_sqmr	154	−9.87	0.29	131	−10.39	0.09	188	−9.88	0.25
bl_bicgstab	395*	−10.09	0.41	†	†	†	433*	−10.04	0.35
bl_bicrstab	356*	−9.96	0.46	†	†	†	417*	−9.71	0.44
bl_idr(4)	269*	−8.57	0.28	†	†	†	334*	−10.10	0.27
bl_gmres(m)	3**	−10.08	24.5	†	†	†	†	†	†

bl_gmres(m). We use the value $m = 80$ for the restart in bl_gmres(m). The experiments have been carried out in double precision floating point arithmetic, without preconditioning, using MATLAB 2014a (64 bit) on PC-Intel(R) Core(TM) i5-3470 CPU 3.20 GHz, 8 GB of RAM.

The different Block algorithms are compared in terms of number of iterations, denoted as *Iters* in the tables, and $\log_{10}$ of the final true relative residual norm defined as $\log_{10}(\|B - AX_{\text{final}}\|_F/\|B\|_F)$, denoted as *TRR*. The iterative solution is started choosing $X_0 = O \in \mathbb{C}^{n\times p}$ as initial guess. The stopping criterion in our runs is the reduction of the norm of the initial Block residual by eight orders of magnitude, i.e., $\|R_m\|_F/\|B\|_F \leq Tol = 10^{-10}$. The right-hand side B is computed by the MATLAB function `rand`. In the tables, the symbol "†" indicates no convergence within n iterations, or n/m cycles for the bl_gmres(m) method.

The first test problems are three matrices extracted from the Matrix Market collection,[2] arising from modeling acoustic scattering problems. They are denoted as `young1c, young2c`, and `young3c`. The results of our experiments are presented in Table 2. The symbol * used for the bl_bicgstab, bl_idr(4), and bl_bicrstab methods indicate that these three methods require no less than two matrix products AV per iteration step. The symbol ** refers to the number of outer iterations in the Block GMRES(m) method, when it can achieve convergence; refer to [15] for details. This notation is used throughout this section.

Table 2 shows the results with nine different Block Krylov solvers. Although the bl_cocg and bl_cocr methods required more *Iters*, they are more competitive than the bl_sqmr method in terms of wall clock time and *TRR* (except the case of `young3c`). Bl_cocr method is more robust than bl_cocg in terms of *Iters*, wall clock time and *TRR*. The bl_cocg_rq and bl_cocr_rq variants are very efficient in terms of *TRR* and wall clock time. The bl_bicgstab, bl_bicrstab, bl_idr(4), and bl_gmres(m) methods cannot solve the test problem (`young3c`), while bl_cocg

[2] http://math.nist.gov/MatrixMarket/matrices.html

Table 3 The numerical results of different iterative solvers for Example 1. The solution time refers to the wall clock time expressed in seconds

	sphere2430			parallelepipede			cube1800		
Method	*Iters*	*TRR*	Time	*Iters*	*TRR*	Time	*Iters*	*TRR*	Time
bl_cocg	189	−10.07	4.16	176	−10.02	2.40	174	−10.21	1.94
bl_cocg_rq	169	−10.00	3.77	156	−10.13	2.13	156	−10.08	1.74
bl_cocr	186	−10.03	4.12	174	−10.02	2.35	169	−10.00	1.84
bl_cocr_rq	166	−10.05	3.77	152	−10.15	2.11	151	−10.09	1.73
bl_sqmr	172	−9.84	4.15	161	−9.91	2.42	159	−9.97	2.11
bl_bicgstab	379*	−10.04	16.5	370*	−10.04	9.94	396*	−10.29	8.42
bl_bicrstab	392*	−9.57	17.3	355*	−9.85	9.98	303*	−8.38	6.70
bl_idr(4)	409*	−9.64	22.1	474*	−10.11	16.5	334*	−9.43	10.2
bl_gmres(m)	2**	−10.07	38.2	2**	−10.04	33.3	2**	−10.09	22.1

and bl_cocr converge rapidly. Due to the long iterative recurrence, the bl_gmres(m) method is typically expensive.

In the second experiment we consider three dense matrices arising from monostatic radar cross-section calculation; they are denoted as `parallelepipede`, `sphere2430`, `cube1800`. These problems are available from our GitHub repository,[3] and we choose $p = 8$. Although rather small, the selected dense problems are representative of realistic radar-cross-section calculation [2]. Larger problems would require a Fortran or C implementation of the solvers and will be considered in a separate study. Numerical results for each test problem are summarized in Table 3.

Table 3 displays the results with again nine different Block Krylov solvers. We can see that the bl_sqmr method requires less *Iters* to converge compared to the bl_cocg and bl_cocr methods. However, it is more expensive in terms of wall clock time except on the `sphere2430` problem. Besides, the true residual norms produced by the bl_sqmr method are larger than those of both bl_cocg and bl_cocr. Furthermore, bl_cocg_rq and bl_cocr_rq are the most effective and promising solvers in terms of *Iters* and wall clock time. Specifically, the bl_cocr_rq method is slightly more efficient than the bl_cocg_rq method in terms of *TRR*.

4 Conclusions

In this paper, a framework for constructing new Block iterative Krylov subspace methods is presented. Two new matrix solvers that can exploit the symmetry of A for solving complex symmetric non-Hermitian linear systems (1) are introduced. Stabilization techniques based on residual orthonormalization strategy are discussed for both methods. The numerical experiments show that the solvers can be viable

[3] https://github.com/Hsien-Ming-Ku/Test_matrices/tree/master/Example2

alternative to standard Krylov subspace methods for solving complex symmetric linear systems with multiple RHSs efficiently. Obviously, for solving realistic electromagnetic problems they both need to be combinated with suitable preconditioners that reflect the symmetry of A; we refer the reader to, e.g., [16–18] for some related studies.

References

1. I.S. Duff, L. Giraud, J. Langou, E. Martin, Using spectral low rank preconditioners for large electromagnetic calculations. Int. J. Numer. Methods Eng. **62**, 416–434 (2005)
2. B. Carpentieri, I.S. Duff, L. Giraud, G. Sylvand, Combining fast multipole techniques and an approximate inverse preconditioner for large electromagnetism calculations. SIAM J. Sci. Comput. **27**, 774–792 (2005)
3. J. Zhang, J. Zhao, A novel class of block methods based on the block AA^T-Lanczos bi-orthogonalization process for matrix equations. Int. J. Comput. Math. **90**, 341–359 (2013)
4. M.H. Gutknecht, Block Krylov space methods for linear systems with multiple right-hand sides: an introduction. in *Modern Mathematical Models, Methods and Algorithms for Real World Systems* ed. by A. H. Siddiqi, I. S. Duff, O. Christensen, (Anamaya Publishers, New Delhi, 2006), pp. 420–447
5. D.P. O'Leary, The block conjugate gradient algorithm and related methods. Linear Algebra Appl. **29**, 293–322 (1980)
6. A. el Guennouni, K. Jbilou, H. Sadok, A block version of BiCGSTAB for linear systems with multiple right-hand sides. Electron. Trans. Numer. Anal. **16**, 129–142 (2003)
7. R.W. Freund, M. Malhotra, A block QMR algorithm for non-Hermitian linear systems with multiple right-hand sides. Linear Algebra Appl. **254**, 119–157 (1997)
8. L. Du, T. Sogabe, B. Yu, Y. Yamamoto, S.-L. Zhang, A block IDR(s) method for nonsymmetric linear systems with multiple right-hand sides. J. Comput. Appl. Math. **235**, 4095–4106 (2011)
9. B. Vital, Etude de quelques méthodes de résolution de problémes linéaires de grande taille sur multiprocesseur, Ph.D. Thesis, Université de Rennes I, Rennes, 1990
10. H. Tadano, T. Sakurai, A block Krylov subspace method for the contour integral method and its application to molecular orbital computations. IPSJ Trans. Adv. Comput. Syst. **2**, 10–18 (2009, in Japanese)
11. H.A. Van der Vorst, J.B.M. Melissen, A Petrov-Galerkin type method for solving $A\boldsymbol{x} = \boldsymbol{b}$, where A is symmetric complex. IEEE Trans. Mag. **26**, 706–708 (1990)
12. T. Sogabe, S.-L. Zhang, A COCR method for solving complex symmetric linear systems. J. Comput. Appl. Math. **199**, 297–303 (2007)
13. A.A. Dubrulle, Retooling the method of block conjugate gradients. Electron. Trans. Numer. Anal. **12**, 216–233 (2001)
14. X.-M. Gu, M. Clemens, T.-Z. Huang, L. Li, The SCBiCG class of algorithms for complex symmetric linear systems with applications in several electromagnetic model problems. Comput. Phys. Commun. **191**, 52–64 (2015)
15. H.-X. Zhong, G. Wu, G. Chen, A flexible and adaptive simpler block GMRES with deflated restarting for linear systems with multiple right-hand sides. J. Comput. Appl. Math. **282**, 139–156 (2015)

16. B. Carpentieri, M. Bollhöfer, Symmetric inverse-based multilevel ILU preconditioning for solving dense complex non-Hermitian systems in electromagnetics. Prog. Electromagn. Res. (PIER) **128**, 55–74 (2012)
17. P.L. Rui, R.S. Chen, Z.H. Fan, D.Z. Ding, Multi-step spectral preconditioner for fast monostatic radar cross-section calculation. Electron. Lett. **43**, 422–423 (2007)
18. B. Carpentieri, I.S. Duff , L. Giraud, M. Magolu monga Made, Sparse symmetric preconditioners for dense linear systems in electromagnetism. Numer. Linear Algebra Appl. **11**, 753–771 (2004)

Part V
Reduced Order Modeling

Model Reduction for Multiscale Lithium-Ion Battery Simulation

Mario Ohlberger, Stephan Rave, and Felix Schindler

Abstract In this contribution we are concerned with efficient model reduction for multiscale problems arising in lithium-ion battery modeling with spatially resolved porous electrodes. We present new results on the application of the reduced basis method to the resulting instationary 3D battery model that involves strong non-linearities due to Buttler-Volmer kinetics. Empirical operator interpolation is used to efficiently deal with this issue. Furthermore, we present the localized reduced basis multiscale method for parabolic problems applied to a thermal model of batteries with resolved porous electrodes. Numerical experiments are given that demonstrate the reduction capabilities of the presented approaches for these real world applications.

1 Introduction

Continuum modeling of batteries results in a reaction-diffusion-transport system of coupled nonlinear partial differential equations in complex multiscale and multi-phase pore structures. In recent contributions [20, 21, 28] three dimensional numerical models have been proposed that resolve the porous electrodes and thus serve as a basis for multiscale modeling as well as for more complex modeling of degradation processes such as Lithium plating. Concerning multiscale modeling in the context of battery simulation, we refer e.g. to [7, 10, 30]. These models result in huge time dependent discrete systems which require enormous computing resources, already for single simulation runs. Parameter studies, design optimization or optimal control, however, require many forward simulation runs with varying material or state parameters and are thus virtually impossible. Hence, model reduction approaches for the resulting parameterized systems are indispensable for such simulation tasks. In this contribution we apply the reduced basis method, that

M. Ohlberger (✉) • S. Rave • F. Schindler
Applied Mathematics Münster, CMTC & Center for Nonlinear Science, University of Münster, Einsteinstr. 62, 48149 Münster, Germany
e-mail: mario.ohlberger@uni-muenster.de; stephan.rave@uni-muenster.de; felix.schindler@uni-muenster.de

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_31

has seen significant advance in recent years. For an overview, we refer to the recent monographs [15, 29] and the tutorial [12].

Concerning model reduction for lithium-ion battery models, we refer to the early work [5] where Galerkin projection into a subspace generated by proper orthogonal decomposition (POD)is used on the basis of the mathematical model proposed in [8]. In [19], the POD approach is used in the context of parameter identification for battery models. Preliminary results concerning model reduction with reduced basis methods can be found in [16, 31] and [27].

In this contribution we focus on two advances in reduced order modeling for batteries. First, in Sect. 2, we present new results concerning nonlinear model reduction for the microscale battery model presented in [20]. The model reduction approach is based on Galerkin projection onto POD spaces, extended to nonlinear problems using empirical operator interpolation [2, 9, 13].

Second, in Sect. 3 we demonstrate the applicability of the localized reduced basis multiscale method (LRBMS) for a thermal model of batteries with resolved porous electrodes. The LRBMS has first been introduced in [1, 18] and further developed in [25, 26]. The later contributions in particular propose a rigorous a posteriori error estimate for the reduced solution with respect to the exact solution for elliptic problems that is localizable and can thus be used to steer an adaptive online enrichment procedure. For an application of the method for more complex problems in the context of two phase flow in porous media we refer to [17]

2 Reduced Basis Methods Applied to Pore-Scale Battery Models

In this section we present first numerical results for the full model order reduction of large 3D pore-scale Li-ion battery models. These results extend our preliminary findings in [27], where we tested the quality of the reduced basis approximation for a small test geometry, towards realistically sized geometries used in real-world simulations, showing the feasibility of our model reduction approach. Before discussing our new results, we will briefly review the battery model under consideration and the basics of the reduced basis methodology.

2.1 A Pore-Scale Lithium-Ion Battery Model

Following [27], we consider a pore-scale battery model based on [20]. The computational domain is divided into five parts: electrolyte, positive/negative electrode, positive/negative current collector (Fig. 1). On each of these subdomains, partial differential equations are given for the Li-ion concentration c and the electrical potential ϕ.

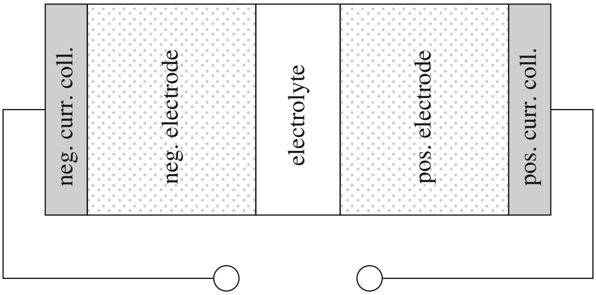


Fig. 1 Schematic overview of the considered battery geometry (note that electrodes have porous structure, pore space is filled with electrolyte)

For the electrolyte we have

$$\frac{\partial c}{\partial t} - \nabla \cdot (D_e \nabla c) = 0, \tag{1}$$

$$-\nabla \cdot \left(\kappa \frac{1-t_+}{F} RT \frac{1}{c} \nabla c - \kappa \nabla \phi \right) = 0, \tag{2}$$

where $D_e = 1.622 \cdot 10^{-6} \frac{\mathrm{cm}^2}{\mathrm{s}}$, $\kappa = 0.02 \frac{\mathrm{s}}{\mathrm{cm}}$, $t_+ = 0.39989$ denote the collective interdiffusion coefficient in the electrolyte, the ion conductivity, and the transference number. $R = 8.314 \frac{\mathrm{J}}{\mathrm{mol\,K}}$, $F = 96487 \frac{\mathrm{As}}{\mathrm{mol}}$ are the universal gas constant and the Faraday constant. We fix the global temperature T to 298K.

In the electrodes, c and ϕ satisfy

$$\frac{\partial c}{\partial t} - \nabla \cdot (D_s \nabla c) = 0, \tag{3}$$

$$-\nabla \cdot (\sigma \nabla \phi) = 0, \tag{4}$$

where $D_s = 10^{-10} \frac{\mathrm{cm}^2}{\mathrm{s}}$ is the ion diffusion coefficient in the electrodes, and $\sigma = 10 \frac{\mathrm{s}}{\mathrm{cm}}$ ($\sigma = 0.38 \frac{\mathrm{s}}{\mathrm{cm}}$) in the negative (positive) electrode denotes the electronic conductivity.

Finally, no Li-ions can enter the current collectors, so $c = 0$ on the whole current collector subdomains. Moreover, ϕ again satisfies

$$-\nabla \cdot (\sigma \nabla \phi) = 0, \tag{5}$$

with $\sigma = 10 \frac{\mathrm{s}}{\mathrm{cm}}$ ($\sigma = 0.38 \frac{\mathrm{s}}{\mathrm{cm}}$) for the negative (positive) current collector.

Note that for this in comparison to [20] slightly simplified model (assuming constant t_+), the Equations (1), (3) are linear and decoupled from the potential equations. However, the coupling between the two variables is established by the interface conditions at the electrode-electrolyte interfaces, where the so-called

Butler-Volmer kinetics are assumed: the electric current (ion flux) j (N) from the electrodes into the electrolyte is given by

$$j = 2k\sqrt{c_e c_s (c_{max} - c_s)} \sinh\left(\frac{\phi_s - \phi_e - U_0(\frac{c_s}{c_{max}})}{2RT} \cdot F \right), \quad N = \frac{j}{F}. \tag{6}$$

Here, $c_{e/s}$ ($\phi_{e/s}$) denotes the Li-ion concentration (electrical potential) at the electrolyte/electrode side of the interface. $c_{max} = 24681 \cdot 10^{-6} \frac{\mathrm{mol}}{\mathrm{cm}^3}$ ($c_{max} = 23671 \cdot 10^{-6} \frac{\mathrm{mol}}{\mathrm{cm}^3}$) denotes the maximum Li-ion concentration in the negative (positive) electrode, and the rate constant k is given by $k = 0.002 \frac{\mathrm{Acm}^{2.5}}{\mathrm{mol}^{1.5}}$ at the negative and by $k = 0.2 \frac{\mathrm{Acm}^{2.5}}{\mathrm{mol}^{1.5}}$ at the positive electrode interface. Finally, the open circuit potential is given by $U_0(s) = (-0.132 + 1.41 \cdot e^{-3.52s})V$ for the negative, and by

$$\begin{aligned} U_0(s) = \Big[& \, 0.0677504 \cdot \tanh(-21.8502 \cdot s + 12.8268) \\ & - 0.105734 \cdot \big((1.00167 - s)^{-0.379571} - 1.576\big) \\ & - 0.045 \cdot e^{-71.69 \cdot s^8} + 0.01 \cdot e^{-200 \cdot (s - 0.19)} + 4.06279 \Big] \cdot V \end{aligned} \tag{7}$$

for the positive electrode.

Given the porous electrode structures, these interface conditions apply to a large surface area, giving this model highly nonlinear dynamics.

Finally, the system is closed by the following boundary conditions: homogeneous Neumann conditions for c at all further inner and external domain boundaries, continuity conditions for ϕ at the current collector-electrode interfaces, homogenous Neumann conditions for ϕ at the current collector-electrolyte interfaces, $\phi \equiv U_0(c(0)/c_{max})$ at the negative current collector boundary, and $-n \cdot \sigma \nabla \phi \equiv \mu$ at the positive current collector boundary.

We consider the fixed charge rate μ as a parameter we want to vary in our numerical experiments.

2.2 Reduced Basis Method and Empirical Interpolation

After cell-centered finite volume discretization of the model on a voxel grid, replacing the numerical fluxes by the Butler-Volmer relations at the electrode-electrolyte interfaces, and backward Euler time discretization, we arrive at nonlinear, discrete equations systems of the form

$$\begin{bmatrix} \frac{1}{\Delta t}(c_\mu^{(t+1)} - c_\mu^{(t)}) \\ 0 \end{bmatrix} + A_\mu \left(\begin{bmatrix} c_\mu^{(t+1)} \\ \phi_\mu^{(t+1)} \end{bmatrix} \right) = 0, \qquad (c_\mu^{(t)}, \phi_\mu^{(t)}) \in V_h \oplus V_h, \tag{8}$$

where A_μ denotes the parametric finite volume space differential operator acting on the finite volume space V_h (see [28] for a detailed derivation). Solving these systems using Newton's method requires many hours for realistic geometries, even when using advanced algebraic multigrid solvers for computing the Newton updates.

Projection-based parametric model reduction methods are based on the idea of finding problem adapted approximation spaces $\tilde{V} \subseteq V_h \oplus V_h$ in which a reduced order solution is obtained by projection of the original equation system:

$$P_{\tilde{V}} \left\{ \begin{bmatrix} \frac{1}{\Delta t}(\tilde{c}_\mu^{(t+1)} - \tilde{c}_\mu^{(t)}) \\ 0 \end{bmatrix} + A_\mu \left(\begin{bmatrix} \tilde{c}_\mu^{(t+1)} \\ \tilde{\phi}_\mu^{(t+1)} \end{bmatrix} \right) \right\} = 0, \quad (\tilde{c}_\mu^{(t)}, \tilde{\phi}_\mu^{(t)}) \in \tilde{V}. \tag{9}$$

Here, $P_{\tilde{V}}$ denotes the orthogonal projection onto $\tilde{V}$. Since the manifold of system states $\{(c_\mu^{(t)}, \phi_\mu^{(t)}) \mid \mu \in [\mu_{min}, \mu_{max}],\ t \in \{0, \dots, T\}\}$ has a low-dimensional parametrization (by $(\mu, t) \in \mathbb{R}^2$), and assuming that this parametrization is sufficiently smooth, there is hope to find low-dimensional approximation spaces $\tilde{V}$ such that the model reduction error between the reduced solutions (9) and the corresponding high-dimensional solutions (8) is very small.

A vast amount of methods for constructing reduced spaces $\tilde{V}$ has been considered in literature. For time-dependent problems, the POD-GREEDY method [11, 14] has shown to produce approximation spaces with quasi-optimal l^∞-in-μ, l^2-in-time reduction error. In our experiments below, we apply a more basic approach by computing a basis for $\tilde{V}$ via PODs of a pre-selected set of solution trajectories of (8). More precisely, we compute separate reduced concentration ($\tilde{V}_c$) and potential ($\tilde{V}_\phi$) spaces and let $\tilde{V} := \tilde{V}_c \oplus \tilde{V}_\phi$. Due to the basic properties of POD, $\tilde{V}_c$, $\tilde{V}_\phi$ are l^2-in-μ, l^2-in-time best-approximation spaces for the considered training set of solutions.

Even though the equation systems (9) are posed on the low-dimensional space $\tilde{V}$, solving (9) requires the evaluation of the projected operator $P_{\tilde{V}} \circ A_\mu$ (and its Jacobian), which in turn makes the computationally expensive evaluation of A_μ on the full finite volume space $V_h \oplus V_h$ necessary. The method of choice to overcome this limitation for nonlinear operators A_μ is empirical operator interpolation: A_μ is replaced by an interpolant $I_M \circ \tilde{A}_{M,\mu} \circ R_{M'}$, where $\tilde{A}_{M,\mu} : \mathbb{R}^{M'} \to \mathbb{R}^M$ is the restriction of A_μ to M appropriately selected degrees of freedom (DOFs), $R_{M'} : V_h \oplus V_h \to \mathbb{R}^{M'}$ is the restriction of the finite volume vectors to the M' DOFs required for the evaluation of $\tilde{A}_{M,\mu}$ and $I_M : \mathbb{R}^M \to V_h \oplus V_h$ is the linear combination with an appropriate interpolation basis (collateral basis). Due to the locality of finite volume operators, M' can be chosen such that $M' \leq C \cdot M$, where C only depends on the maximum number of neighboring cells in the given mesh. The interpolation DOFs and the associated collateral basis are obtained from solution snapshot data using the EI-GREEDY algorithm [9, 13].

A direct application of this approach to A_μ would not be successful, however: since the collateral basis is contained in the linear span of operator evaluations on solution trajectories, the ϕ-parts of the collateral basis vectors would, according to (8), completely vanish. Therefore, we first decompose A_μ as $A_\mu = A^{(const)} + \mu \cdot A^{(bnd)} + A^{(lin)} + A^{(1/c)} + A^{(bv)}$, where $A^{(1/c)}$, $A^{(bv)}$ are the nonlinear operators corresponding to $-\nabla \cdot \kappa \frac{1-t_+}{F} RT \frac{1}{c} \nabla c$ and the Butler-Volmer interfaces, $A^{(const)}$ ($A^{(bnd)}$) is the constant (parametric) part of A_μ corresponding to the boundary conditions, and $A^{(lin)}$ is the remaining linear part of A_μ. We then apply empirical operator interpolation separately to $A^{(1/c)}$ and $A^{(bv)}$. With $T[\tilde{c}_\mu^{(t)}](\tilde{c}, \tilde{\phi}) := (1/\Delta t \cdot (\tilde{c} - \tilde{c}_\mu^{(t)}),\ 0)$, we arrive at the fully reduced systems

$$\begin{aligned} \Big\{ T[\tilde{c}_\mu^{(t)}] &+ P_{\tilde{V}} \circ A^{(const)} + \mu \cdot P_{\tilde{V}} \circ A^{(bnd)} + P_{\tilde{V}} \circ A^{(lin)} \\ &+ \{P_{\tilde{V}} \circ I^{(1/c)}_{M^{(1/c)}}\} \circ \tilde{A}^{(1/c)}_{M^{(1/c)},\mu} \circ R^{(1/c)}_{M'^{(1/c)}} \\ &+ \{P_{\tilde{V}} \circ I^{(bv)}_{M^{(bv)}}\} \circ \tilde{A}^{(bv)}_{M^{(bv)},\mu} \circ R^{(bv)}_{M'^{(bv)}} \Big\} \left(\begin{bmatrix} \tilde{c}_\mu^{(t+1)} \\ \tilde{\phi}_\mu^{(t+1)} \end{bmatrix} \right) = 0. \end{aligned} \tag{10}$$

After pre-computation of the linear maps $P_{\tilde{V}} \circ A^{(bnd)}$, $P_{\tilde{V}} \circ A^{(lin)}$, $P_{\tilde{V}} \circ I^{(1/c)}_{M^{(1/c)}}$, $R^{(1/c)}_{M'^{(1/c)}}$, $P_{\tilde{V}} \circ I^{(bv)}_{M^{(bv)}}$, $R^{(bv)}_{M'^{(bv)}}$ and of the constant map $P_{\tilde{V}} \circ A^{(const)}$ w.r.t. to a basis of $\tilde{V}$, (10) can be solved quickly and independent of the dimension of V_h.

2.3 Numerical Experiments

We consider two different test cases: a small test geometry (Fig. 2) which still exhibits the most important properties of a real battery geometry, and a large, fully resolved geometry (Fig. 3) useable for real-world simulations. In both cases, the initial Li$^+$ concentration c_0 was set to $c_0 \equiv 2639 \cdot 10^{-6} \frac{\text{mol}}{\text{cm}^3}$ ($c_0 \equiv 20574 \cdot 10^{-6} \frac{\text{mol}}{\text{cm}^3}$) for the positive (negative) electrode and to $c_0 \equiv 1200 \cdot 10^{-6} \frac{\text{mol}}{\text{cm}^3}$ in the electrolyte. The model was simulated on a $T = 2000$s ($T = 1600$s) time interval for the small (large) geometry, with a time step size of $\Delta t = 20$s. The charge rate μ was for each simulation chosen as a constant from the interval $\left[0.00012 \frac{\text{A}}{\text{cm}^2}, 0.0012 \frac{\text{A}}{\text{cm}^2}\right]$ for the small and from the interval $\left[0.000318 \frac{\text{A}}{\text{cm}^2}, 0.00318 \frac{\text{A}}{\text{cm}^2}\right]$ for the large geometry.

To generate the reduced space $\tilde{V}$, we computed solution snapshots on training sets $\mathscr{S}_{train}$ of equidistant parameters. For the small geometry we chose $\#\mathscr{S}_{train} = 20$, whereas for the large geometry we only selected the lower and upper boundary of the considered parameter domain, i.e. $\#\mathscr{S}_{train} = 2$. For the generation of the empirical interpolation data using the EI-Greedy algorithm, we additionally included the evaluations of $A_\mu^{(1/c)}$ and $A_\mu^{(bv)}$ on all intermediate Newton stages of the selected solution trajectories.

As a measure for the model reduction error we consider the relative l^∞-in-μ, l^∞-in-time error given by

$$\max_{\mu \in \mathscr{S}_{test}} \max_{t \in \{0,1,\ldots T/\Delta t\}} \frac{\|u_\mu^{(t)} - \tilde{u}_\mu^{(t)}\|}{\max_{t \in \{0,1,\ldots T/\Delta t\}} \|u_\mu^{(t)}\|}, \tag{11}$$

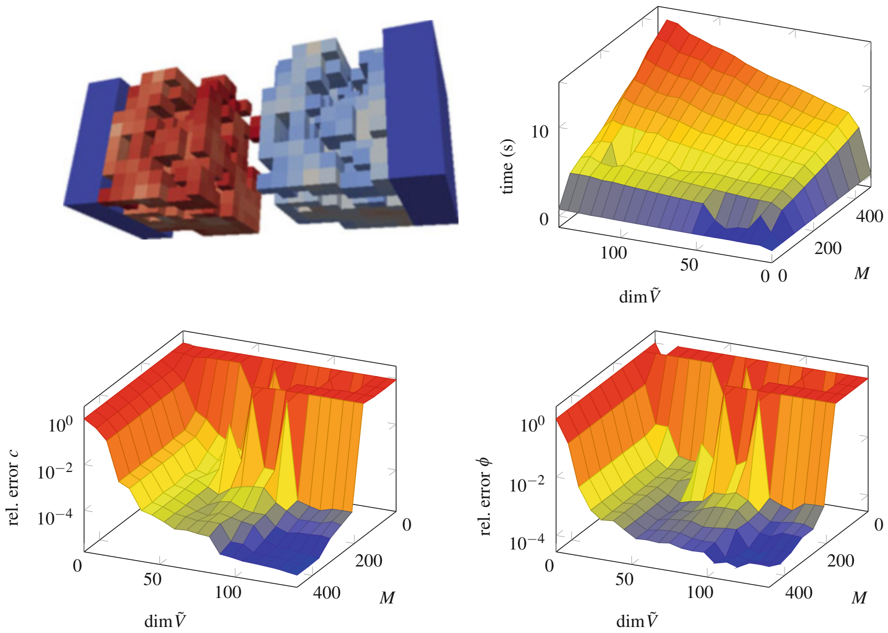


Fig. 2 *Top left*: small porous battery geometry used in numerical experiments. Size: 104 × 40 × 40 μm, 4.600 DOFs, coloring indicates Li$^+$ concentration at end of simulation, electrolyte not depicted. *Top right*: average solution time in seconds vs. dimension of reduced space $\tilde{V}$ and number of interpolation points ($M := M^{(1/c)} + M^{(bv)}$). Relative model reduction errors (11) for concentration (*middle*) and potential (*bottom*) variable vs. dimension of reduced space and number of interpolation points. A training set of 20 equidistant parameters was used for the generation of $\tilde{V}$ and the interpolation data, $\#\mathscr{S}_{test} = 20$

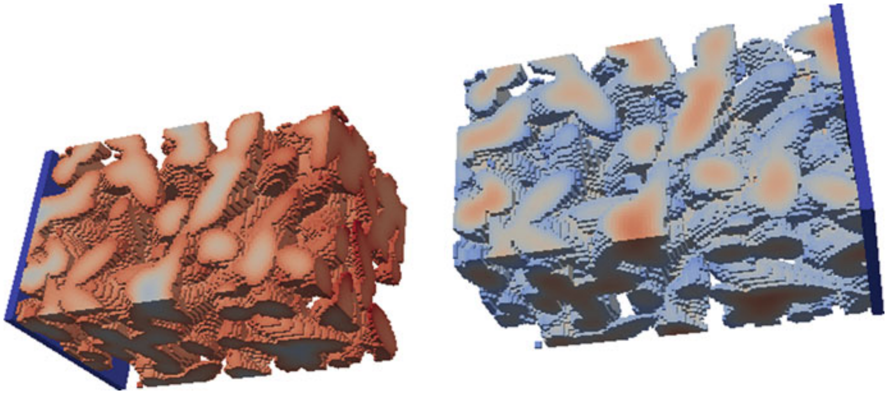


Fig. 3 Porous battery geometry used in the numerical experiments. Size: 246 × 60 × 60 μm, 1.749.600 DOFs, coloring indicates Li$^+$ concentration at end of simulation, electrolyte not depicted

where u ($\tilde{u}$) is the concentration or potential part of the (reduced) solution and $\mathscr{S}_{test}$ denotes a random set of test parameters.

All simulations of the high-dimensional model have been performed with the battery simulation software `BEST` [21], which has been integrated with our model order reduction library `pyMOR` [22, 27]. The experiments were conducted as single-threaded processes on a dual socket compute server equipped with two Intel Xeon E5-2698 v3 CPUs with 16 cores running at 2.30 GHz each and 256 GB of memory available.

For the small test geometry, we observe a rapid decay of the model reduction error for both the concentration and the potential variable (Fig. 2). As usual for empirical operator interpolation, we see that the number of interpolation points has to be increased for larger reduced space dimensions in order to ensure stability of the reduced model. Doing so, we obtain relative reduction errors as small as 10^{-4} with simulation times of less than 15 s.

Since we only selected 2 solution trajectories for the generation of the reduced model for the large geometry, we cannot expect such small model reduction errors over the whole parameter domain. In fact, the error stagnates already for relatively small reduced space dimensions (Table 1). Nevertheless, we easily achieve errors of less than 1 % for a simulation time of 80 s. With an average solution time for the high-dimensional model of over 6 h, we achieve at this error a speedup factor of 285.

Note that the solution time of the reduced model is still significantly larger than for the small geometry. This can be attributed to the fact that the localized evaluation of $A_\mu^{(1/c)}$, $A_\mu^{(bv)}$ has been only partially implemented in `BEST` and still requires operations on high-dimensional data structures. After the implementation of localized operator evaluation in `BEST` has been finalized, we expect even shorter simulation times.

Table 1 Relative model reduction errors (11) and reduced simulation times for the large battery geometry (Fig. 3). 188 interpolation points, average time for solution of the high-dimensional model: 22,979s, $\#\mathscr{S}_{test} = 10$

$\dim \widetilde{V}$	11	21	30	40
Rel. error c	$9.26 \cdot 10^{-3}$	$3.96 \cdot 10^{-3}$	$3.05 \cdot 10^{-3}$	$2.93 \cdot 10^{-3}$
Rel. error ϕ	$2.07 \cdot 10^{-3}$	$1.50 \cdot 10^{-3}$	$1.46 \cdot 10^{-3}$	$1.26 \cdot 10^{-3}$
Time (s)	82	81	79	81
Speedup	279	285	290	283

3 Localized Reduced Basis Multiscale Approximation of Heat Conduction

The microscale battery model in Sect. 2 is considered under the assumption of constant global temperature T. In general, it is desirable to couple this model with a spatially resolved model for the temperature distributions within the battery. For the model reduction of such heat conduction in porous electrodes we present a first application of the localized reduced basis multiscale Method (LRBMS) for parabolic PDEs.

In this first step we consider the simulation and model reduction of heat conduction separately from what is presented in Sect. 2 as a basis for a coupled simulation and model reduction in future work.

For an introduction of the LRBMS for elliptic parameterized multiscale problems and recent results concerning localized a posteriori error estimation and online enrichment, we refer to [26].

3.1 A Battery: Heat Conduction Model with Resolved Electrode Geometry

We consider here the same spatially resolved 3D pore-scale battery geometry (cf. Fig. 3) as in Sect. 2, where the computational domain is composed of five materials which are of interest for thermal modeling, that is: electrolyte, positive/negative electrode and positive/negative current collectors, each with possibly different thermal conductivities.

As a simplified model for heat conductivity within a battery with spatially resolved electrodes, we consider a parabolic PDE for the temperature T of the form

$$\frac{\partial T}{\partial t} - \nabla \cdot \big(D \, \nabla T\big) = Q, \tag{12}$$

together with suitable initial and boundary conditions. Here D denotes the space-dependent conductivity tensor, which is material specific and thus takes different values in the current collectors, the porous electrodes, the separator, and the electrolyte. Hence, D inherits the highly heterogeneous structure of the porous electrodes and thus has an intrinsic multiscale character. In general, Q collects all heat generating sources, such as heat generation due to electrochemical reaction, reversible heat and ohmic heat, each of which may in turn depend on the Li-ion concentration and the electric potential and thus vary in space and time. These sources arise in particular due to the electrochemical reaction at the interface between the electrodes and the electrolyte and it is thus desirable to consider the full 3D pore-scale battery model in order to get an insight into possible variations of the temperature within the battery. We refer, e.g. to[5, 6] for a more detailed derivation of an energy balance equation for Lithium-Ion batteries and corresponding simulation schemes.

Depending on the study in question, any of the sources, the thermal conductivity or the initial or boundary values may depend on a low-dimensional parameter vector μ.

3.2 Localization of Reduced Basis Methods: LRBMS

As a first step towards a realistic model we allow for parametric thermal conductivities and presume stationary sources and boundary values. Thus, a (spatial) discretization of (12) by a suitable discretization scheme (such as finite volumes or continuous or discontinuous Galerkin finite elements) and a backward Euler time-discretization yield a set of linear equations of the form,

$$\tfrac{1}{\Delta t} M_h\big(T^{(t+1)} - T^{(t)}\big) + B_{h,\mu}\, T^{(t+1)} = Q_h, \qquad T^{(t+1)} \in V_h, \tag{13}$$

to be solved in each time step, where M_h and $B_{h,\mu}$ denote the discrete L^2-inner product and parametric space differential operators induced by the spatial discretization, respectively, which act on the corresponding high-dimensional discrete space V_h. In addition, Q_h denotes the discrete representation of the source and boundary values.

To obtain a reduced order model for the discrete heat conduction model (13), we proceed in an analog way, as described in Sect. 2 above, by a Galerkin projection onto a problem adapted reduced approximation space $\tilde{V} \subset V_h$. Once $\tilde{V}$ is given, we obtain the set of reduced equations for each time step:

$$\tfrac{1}{\Delta t} \tilde{M}\big(\tilde{T}^{(t+1)} - \tilde{T}^{(t)}\big) + \tilde{B}_{\mu}\, \tilde{T}^{(t+1)} = \tilde{Q}, \qquad \tilde{T}^{(t+1)} \in \tilde{V}, \tag{14}$$

where $\tilde{M}$, $\tilde{B}_\mu$ and $\tilde{Q}$ denote the reduced operators and functionals, respectively, acting on the low-dimensional reduced space $\tilde{V}$. Since all operators and functionals arising in (14) are affinely decomposable with respect to the low-dimensional parameter vector μ (given for instance the thermal conductivity as in Sect. 3.3) and linear with respect to $\tilde{V}$, we can precompute their respective evaluations in a computationally expensive offline step, e.g., by $\tilde{M} = \underline{P_{\tilde{V}}}^{\perp} \underline{M_h}\ \underline{P_{\tilde{V}}}$, where $\underline{M_h}$ and $\underline{P_{\tilde{V}}}$, respectively, denote the matrix representations of M_h and of the orthogonal projection $P_{\tilde{V}} : V_h \to \tilde{V}$ with respect to the basis of V_h. Online, for each new input parameter μ, we can then quickly solve the reduced low-dimensional problem (14) to obtain a low-dimensional representation of the temperature $\tilde{T}$, which can be post-processed to obtain the original temperature T, if required, or a derived quantity of interest.

As mentioned above, the problem adapted reduced space $\tilde{V}$ can be adaptively generated by an iterative POD-GREEDY procedure [14]: in each step of the greedy algorithm, given an error estimate on the model reduction error, a full high-dimensional solution trajectory for the hitherto worst-approximated parameter is computed and the most dominant POD modes of the projection error of this trajectory are added to the reduced basis spanning $\tilde{V}$.

This procedure has been shown to produce quasi-optimal low-dimensional reduced order models which successfully capture the dynamics of the original high-dimensional model [11]. However, in the context of multiscale phenomena or highly resolved geometries, such as the porous structures within a Li-ion battery, the computational cost required to generate the reduced model can become unbearably large, even given modern computing hardware.

As a remedy, the localized reduced basis multiscale method has been introduced for stationary elliptic multiscale problems [1, 18] to lower the computational burden of traditional RB methods by generating several local reduced bases associated with a partitioning of the computational domain. The local quantities associated with these individual subdomains can be projected independently in parallel. In [25, 26], the LRBMS was extended to additionally account for the discretization error and to allow for an adaptive enrichment of the local reduced approximation spaces, which may even eliminate the need for global solution snapshots at all.

In this contribution, we demonstrate a first application of the LRBMS to parabolic multiscale problems, such as spatially resolved heat conduction in a Lithium-Ion battery. We therefore discretize (12) locally by a standard finite element or discontinuous Galerkin scheme independently in each subdomain of a given partitioning of the computational domain and couple the arising local operators, products and functionals along these subdomains by symmetric weighted interior penalty discontinuous Galerkin fluxes (cf. [26] and the references therein). We use the resulting discretization to compute global solution snapshots during the greedy algorithm, as detailed above. However, instead of a single reduced basis with global support, we iteratively generate local reduced bases on each subdomain by localizing the solution trajectories with respect to each subdomain and by carrying out local PODs for further localized compression in a post-processing step.

The resulting reduced space is then given as the direct sum of the local reduced approximation spaces spanned by these local reduced bases. Accordingly, we obtain the reduced problem (14) by local Galerkin projections of the local operators and functionals and coupling operators associated with each subdomain and its neighbor, yielding sparse reduced quantities.

3.3 Numerical Experiments

To demonstrate the applicability of the LRBMS we conduct an experiment on the same geometry used in the larger experiment in Sect. 2.3 (compare Fig. 3). For the thermal conductivities we choose constant values within each material (the positive/negative electrode and the positive/negative current collectors), as reported in [6, 4th column of Table 4]. Within the electrolyte we allow to vary the constant thermal conductivity within the range $\mu \in [0.1; 10]$. We pose homogeneous Dirichlet boundary values at the current collectors and homogeneous Neumann boundary values elsewhere and start the simulations with an initial temperature of 0K, using ten time steps to reach the final time 10^{-3}. For the heat source we set $Q = 10^3$ within the electrodes and $Q = 0$ within the current collectors and the electrolyte. While this is not necessarily a physically meaningful setup, it inherits the computational challenges of a realistic model, namely a highly resolved geometry, discontinuous thermal conductivities depending on the materials and heat sources which align with the geometry of the different materials.

We triangulate the computational domain with 5,313,600 simplexes and compare the LRBMS using 8×2×2 subdomains to a standard RB method (which corresponds to choosing one subdomain). Within each subdomain, we use the same SWIPDG discretization as for the coupling, thus yielding comparable discretizations with 21,254,400 degrees of freedom in both approaches. As an estimate on the model reduction error we use the true L^∞-in-time, H^1-in-space error.

The discretization is implemented within the `DUNE` numerics environment [3, 4], centered around `dune-gdt` [23]: the `dune-stuff` [24] module provides classes for vectors, matrices and linear solvers (for instance the `bicgstab.amg.ilu0` solver used in these experiments), `dune-gdt` provides the discretization building blocks (such as discrete function spaces, operators, products and functionals), and `dune-hdd`[1] provides parametric discretizations compatible with `pyMOR` [22]. Finally, `dune-pymor`[2] is used, as it provides the `Python`-bindings and wrappers to integrate the `DUNE`-code with our model reduction framework `pyMOR`. The experiments were conducted on the same compute server as described in Sect. 2.3.

As we observe from Fig. 4, both the LRBMS and the standard RB method show comparable exponential error decay. In general, the quality of the reduced spaces generated by the LRBMS is slightly better, while requiring less detailed solution snapshots to reach the same target error.

As can be seen from Table 2, the POD-Greedy basis generation using 32 subdomains is slightly faster than the basis generation using a single subdomain. However, since the experiments were conducted as single-threaded processes and since the LRBMS allows for parallel local PODs and parallel local reduced basis

[1] https://github.com/pymor/dune-hdd

[2] https://github.com/pymor/dune-pymor

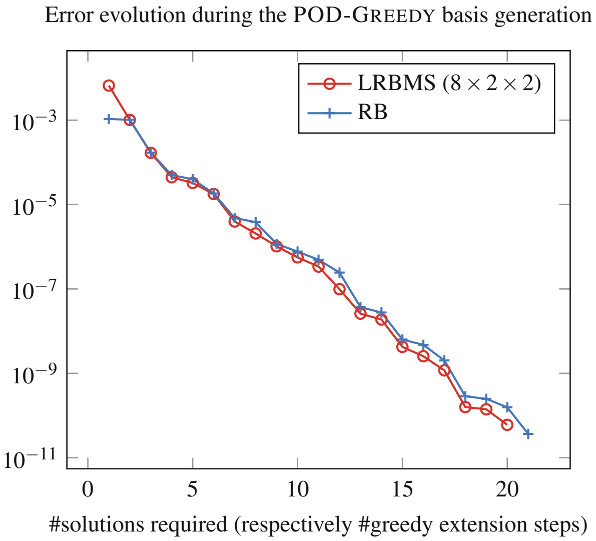


Fig. 4 Error evolution during the POD-GREEDY basis generation to reach a target absolute error of 10^{-10} for the numerical experiment from Sect. 3.3. Depicted is the L^∞-in-μ, L^∞-in-t, and H^1-in-space error over the set of five equidistant training samples in [0.1; 10]

Table 2 Comparison of runtimes of the experiments from Sect. 3.3. Setup time includes grid generation, subdomain partitioning and assembly of operators, products and functionals. POD-GREEDY time includes error estimation, generation of the reduced basis and the reduced basis projection. The average time to solve the detailed problem is 2 h 28 min 5 s

	Setup	POD-GREEDY	Reduced basis size	Solution time
RB	26 min 47 s	14 h 41 min 52 s	21	35 s
LRBMS	36 min 7 s	14 h 34 min 39 s	32 × 20	35 s

projections, the basis generation time of the LRBMS can be further accelerated significantly.

4 Conclusion

In this contribution we have demonstrated the efficient applicability of recent model reduction approaches, such as the POD-GREEDY reduced basis method, the empirical operator interpolation, and the localized reduced basis multiscale method (LRBMS) for efficient simulation of real world problems, such as 3D spatially resolved heterogeneous Lithium-Ion battery models. The demonstrated model reduction approaches are realized within our model order reduction library `pyMOR` [22, 27] with bindings, both to the battery simulation software `BEST` [21], and the general purpose Distributed and Unified Numerics Environment `DUNE`

[3, 4], employing the dune-gdt, dune-stuff, and dune-hdd discretization and solver backends. Speedup factors of about 285 were obtained for the full strongly non-linear battery model in Sect. 2 using the reduced basis method with empirical operator interpolation [9], andaround 253 for the linear parabolic heat conduction model in Sect. 3 using a parabolic extension of the localized reduced basis multiscale method [26].

Acknowledgements The authors thank Sebastian Schmidt from Fraunhofer ITWM Kaiserslautern for the close and fruitful collaboration within the BMBF-project MULTIBAT towards integration of BEST with pyMOR. This work has been supported by the German Federal Ministry of Education and Research (BMBF) under contract number 05M13PMA.

References

1. F. Albrecht, B. Haasdonk, S. Kaulmann, M. Ohlberger, The localized reduced basis multiscale method, in *ALGORITMY 2012 – Proceedings of contributed papers and posters*, ed. by A. Handlovicova, Z. Minarechova, D. Cevcovic, vol. 1 (Slovak University of Technology in Bratislava, Publishing House of STU, 2012) pp. 393–403
2. M. Barrault, Y. Maday, N. Nguyen, A. Patera, An “empirical interpolation” method: application to efficient reduced-basis discretization of partial differential equations. Comptes Rendus de l’Académie des Sciences, Series I **339**, 667–672 (2004)
3. P. Bastian, M. Blatt, A. Dedner, C. Engwer, R. Klöfkorn, R. Kornhuber, M. Ohlberger, O. Sander, A generic grid interface for parallel and adaptive scientific computing. II. Implementation and tests in DUNE. Computing **82**(2–3), 121–138 (2008)
4. P. Bastian, M. Blatt, A. Dedner, C. Engwer, R. Klöfkorn, M. Ohlberger, O. Sander, A generic grid interface for parallel and adaptive scientific computing. I. Abstract framework. Computing **82**(2–3), 103–119 (2008)
5. L. Cai, R. White, Reduction of model order based on proper orthogonal decomposition for lithium-ion battery simulations. J. Electrochem. Soc. **156**(3), A154–A161, cited By 67 (2009)
6. S.C. Chen, C.C. Wan, Y.Y. Wang, Thermal analysis of lithium-ion batteries. J. Power Sources **140**, 111–124 (2005)
7. F. Ciucci, W. Lai, Derivation of micro/macro lithium battery models from homogenization. Transp. Porous Media **88**(2), 249–270 (2011)
8. M. Doyle, T. Fuller, J. Newman, Modeling of galvanostatic charge and discharge of the lithium/polymer/insertion cell. J. Electrochem. Soc. **140**(6), 1526–1533, cited By 789 (1993)
9. M. Drohmann, B. Haasdonk, M. Ohlberger, Reduced basis approximation for nonlinear parametrized evolution equations based on empirical operator interpolation. SIAM J. Sci. Comput. **34**(2), A937–A969 (2012)
10. S. Golmon, K. Maute, M.L. Dunn, Multiscale design optimization of lithium ion batteries using adjoint sensitivity analysis. Internat. J. Numer. Methods Eng. **92**(5), 475–494 (2012)
11. B. Haasdonk, Convergence rates of the POD-Greedy method. M2AN Math. Model. Numer. Anal. **47**, 859–873 (2013)
12. B. Haasdonk, Reduced basis methods for parametrized PDEs – A tutorial introduction for stationary and instationary problems, Technical report, 2014, Chapter to appear in P. Benner, A. Cohen, M. Ohlberger and K. Willcox: “Model Reduction and Approximation: Theory and Algorithms”, SIAM
13. B. Haasdonk, M. Ohlberger, G. Rozza, A reduced basis method for evolution schemes with parameter-dependent explicit operators. Electron. Trans. Numer. Anal. **32**, 145–161 (2008)

14. B. Haasdonk, M. Ohlberger, Reduced basis method for finite volume approximations of parametrized linear evolution equations. M2AN Math. Model. Numer. Anal. **42**(2), 277–302 (2008)
15. J.S. Hesthaven, G. Rozza, B. Stamm, SpringerBriefs in Mathematics. (Springer International Publishing, Heidelberg, 2016)
16. O. Iliev, A. Latz, J. Zausch, S. Zhang, On some model reduction approaches for simulations of processes in Li-ion battery, in *Proceedings of Algoritmy 2012, Conference on Scientific Computing* (Slovak University of Technology in Bratislava, Vysoké Tatry, Podbanské, 2012), pp. 161–171
17. S. Kaulmann, B. Flemisch, B. Haasdonk, K.-A. Lie, M. Ohlberger, The localized reduced basis multiscale method for two-phase flows in porous media. Internat. J. Numer. Methods Eng. **102**(5), 1018–1040 (2015)
18. S. Kaulmann, M. Ohlberger, B. Haasdonk, A new local reduced basis discontinuous Galerkin approach for heterogeneous multiscale problems. C. R. Math. Acad. Sci. Paris **349**(23–24), 1233–1238 (2011)
19. O. Lass, S. Volkwein, Parameter identification for nonlinear elliptic-parabolic systems with application in lithium-ion battery modeling. Comput. Optim. Appl. **62**(1), 217–239 (2015)
20. A. Latz, J. Zausch, Thermodynamic consistent transport theory of Li-ion batteries. J. Power Sources **196**(6), 3296–3302 (2011)
21. G.B. Less, J.H. Seo, S. Han, A.M. Sastry, J. Zausch, A. Latz, S. Schmidt, C. Wieser, D. Kehrwald, S. Fell, Micro-scale modeling of Li-ion batteries: parameterization and validation. J. Electrochem. Soc. **159**(6), A697 (2012)
22. R. Milk, S. Rave, F. Schindler, pyMOR - generic algorithms and interfaces for model order reduction (2015), arXiv e-prints **1506.07094**, http://arxiv.org/abs/1506.07094.
23. R. Milk, F. Schindler, dune-gdt (2015), (http://dx.doi.org/10.5281/zenodo.35389)
24. R. Milk, F. Schindler, dune-stuff (2015), (http://dx.doi.org/10.5281/zenodo.35390)
25. M. Ohlberger, F. Schindler, A-posteriori error estimates for the localized reduced basis multiscale method, in *Finite Volumes for Complex Applications VII-Methods and Theoretical Aspects*, ed. by (J. Fuhrmann, M. Ohlberger, C. Rohde. Springer Proceedings in Mathematics & Statistics, vol. 77 (Springer International Publishing, Berlin, 2014), pp. 421–429
26. M. Ohlberger, F. Schindler, Error control for the localized reduced basis multi-scale method with adaptive on-line enrichment. SIAM J. Sci. Comput. **37**(6), A2865–A2895 (2015)
27. M. Ohlberger, S. Rave, S. Schmidt, S. Zhang, A model reduction framework for efficient simulation of Li-ion batteries, in *Finite Volumes for Complex Applications. VII. Elliptic, Parabolic and Hyperbolic Problems*. Springer Proceedings in Mathematics & Statistics, vol. 78 (Springer, Cham, 2014), pp. 695–702
28. P. Popov, Y. Vutov, S. Margenov, O. Iliev, Finite volume discretization of equations describing nonlinear diffusion in Li-ion batteries, in *Numerical Methods and Applications*, ed. by I. Dimov, S. Dimova, N. Kolkovska. Lecture Notes in Computer Science, vol. 6046 (Springer, Berlin/Heidelberg, 2011), pp. 338–346
29. A. Quarteroni, A. Manzoni, F. Negri, *Reduced Basis Methods for Partial Differential Equations: An Introduction.* Unitext, vol. 92. (Springer, Cham, 2016). La Matematica per il 3+2
30. V. Taralova, Upscaling approaches for nonlinear processes in lithium-ion batteries, Ph.D. thesis, Kaiserslautern, Technische Universität Kaiserslautern, 2015, pp. VII, 224
31. A. Wesche, S. Volkwein, The reduced basis method applied to transport equations of a lithium-ion battery. COMPEL: Int. J. Comput. Math. Electr. Electron. Eng. **32**, 1760–1772 (2013)

Multiscale Model Reduction Methods for Flow in Heterogeneous Porous Media

Assyr Abdulle and Ondrej Budáč

Abstract In this paper we provide a general framework for model reduction methods applied to fluid flow in porous media. Using reduced basis and numerical homogenization techniques we show that the complexity of the numerical approximation of Stokes flow in heterogeneous media can be drastically reduced. The use of such a computational framework is illustrated at several model problems such as two and three scale porous media.

1 Introduction

Fluid flow in porous media is an important and extensively studied process in various applications. Depending on the application, different model and description of a porous medium are used. One of the oldest models is the Darcy equation, which is an elliptic partial differential equation (PDE), that describes an effective fluid flow and pressure in a porous medium [12]. The porous structure, whose geometry is not present in the Darcy model, is accounted for in a permeability tensor. A more precise description is obtained by considering the porous structure explicitly. Knowledge of the geometry of the porous material allows to use a standard model of a fluid flow around obstacles. One can use the Navier-Stokes equation but also the Stokes equation, since the Reynolds number in porous media is often very small.

Let us briefly compare the aforementioned Darcy and fine scale Stokes models. To apply the Darcy model, the permeability tensor of the material is needed. It may be known for standard materials, it can sometimes be obtained experimentally, or, as we present below, it can be computed from the fine scale material structure. The fine scale Stokes approach does not need any effective material property but the computational effort of a direct numerical implementation scales with ratio between the macroscopic domain of interest and the size (typically micrometer) of the pore structure. Hence, this approach is unfeasible for fine porous structures since the number of degrees of freedom is prohibitive.

A. Abdulle (✉) • O. Budáč
École Polytechnique Fédérale de Lausanne, ANMC, CH-1015 Lausanne, Switzerland
e-mail: assyr.abdulle@epfl.ch; ondrej.budac@epfl.ch

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_32

Numerical methods that combine both models and bridge the Darcy and the Stokes scale have been developed, see [2, 8, 10] and the references therein. The Darcy model is used on the macro scale and the effective permeability is upscaled from localized fine scale Stokes computations. This upscaling is based on the homogenization theory [7, 16, 17], which established that a suitable upscaling of the Stokes model leads to the Darcy model. As an example of a numerical realization of this mathematical upscaling procedure we briefly describe the *Darcy-Stokes finite element heterogeneous multiscale method* (DS-FE-HMM) that was introduced in [2]. The *finite element method* (FEM) with numerical quadrature is applied at the macro scale to discretize the Darcy equation and the permeability tensor is recovered at suitable quadrature points. Around every quadrature point we sample the microstructure of the material and solve a Stokes micro problem in a micro domain. The velocity solutions of the micro problems are then avereged to obtain an approximation of the effective permeability that is in turn used to solve the macroscopic Darcy problem. This approach avoids discretization of the whole fine scale porous structure of the material and only zooms on the microstructure where needed.

Most of the multiscale numerical methods for fluid flow in porous media are indeed two-scale, since they consider only the macroscopic (Darcy) scale and the microscopic (Stokes) scale. In practice, however, there are interesting physical processes at more than two scales, for example manufacturing of textile microstructures [13]. Such materials do not fit well into the two-scale setting and modeling that goes beyond two scales is needed. We mention for example [1, 14] where multiscale methods for $n-$scale model (all of which of Darcy type) have been developed. For simplicity, we consider here three-scale models but with different physical model at each scale. The macroscopic description is again the Darcy model with a permeability recovered from a mesoscopic scale, where the fluid flow is described by the Stokes-Brinkman equation. The structure of the porous parts of the mesoscopic domains is described at an even finer scale, the microscopic scale, where the Stokes model is used. We note that the Stokes-Brinkman equation provides a simple coupling of the Stokes equation in the mesoscopic fluid part and the Darcy equation in the mesoscopic porous part. The permeability in the mesoscopic porous part is upscaled from the Stokes micro problems.

Both two-scale and three-scale numerical methods are computationally intensive since we compute a large number of local meso or micro problems ("the cell problems") that are used to upscale the permeability tensor. Errors that are committed by numerical approximation on all scales need to be balanced to obtain an efficient method. While the time cost of such coupled micro-meso-macro multiscale methods does not depend on the pore sizes, it still grows quickly while refining the macroscopic domain. One approach to reduce the computational time cost is to adaptively control the refinement on each scale, which was successfully applied in the two-scale settings [2]. Further reductions are possible by exploiting redundancy in cell problems. Model reduction techniques such as the *reduced basis* (RB) method [3] can be applied to select only the most significant cell problems which can lead to a speed up of orders of magnitude [4].

In this paper we review two-scale and three-scale porous media and multiscale model reduction methods for fluid flow in such media. The cell problems (micro and meso) can be paramterized and formulated in a common framework that is suitable for the RB method. The main element of the RB method is an affine decomposition of the parametric problem, which needs to be provided for the cell problems. The *empirical interpolation method* (EIM) [9] is an important tool to obtain such an affine decomposition of the meso scale [5].

This paper is structured as follows. In Sect. 2 we introduce the two- and three-scale porous media and the corresponding flow models. Numerical homogenization methods for such models are described in Sect. 3 and the combination with model order reduction techniques is presented in Sect. 4. Numerical experiments that illustrate the behavior of the multiscale model reduction methods are provided in Sect. 5.

2 Multiscale Porous Media and Flow Models

Let $d \in \{2,3\}$ and $\Omega \subset \mathbb{R}^d$ be a connected bounded domain in which we consider a porous medium represented by a fluid subset $\Omega_\varepsilon \subset \Omega$, where $\varepsilon > 0$ denotes the microscopic feature scale. Fluid flow in Ω_ε can be modeled by the Stokes equation: find a velocity field $\mathbf{u}^\varepsilon$ and a pressure p^ε such that

$$\begin{aligned} -\Delta \mathbf{u}^\varepsilon + \nabla p^\varepsilon &= \mathbf{f} \quad \text{in } \Omega_\varepsilon, \\ \operatorname{div} \mathbf{u}^\varepsilon &= 0 \quad \text{in } \Omega_\varepsilon, \\ \mathbf{u}^\varepsilon &= 0 \quad \text{on } \partial\Omega_\varepsilon, \end{aligned} \tag{1}$$

where $\mathbf{f}$ is a given force field. For $\varepsilon \ll \operatorname{diam}(\Omega)$ the geometry of Ω_ε is too complex, which makes its meshing and direct numerical solution to (1) prohibitive. Instead, we examine the limit behavior of the solution $(\mathbf{u}^\varepsilon, p^\varepsilon)$ for $\varepsilon \to 0$, which is studied by the homogenization theory. An effective limit solution can be derived in various situations, in particular for periodic porous media [6, 16, 17] and locally periodic porous media [2, 11], as follows. First, we extend the solution $(\mathbf{u}^\varepsilon, p^\varepsilon)$ from Ω_ε to Ω and denote it $(\mathbf{U}^\varepsilon, P^\varepsilon)$. Second, it can be shown that there exist a homogenized pressure p^0 and a homogenized velocity field $\mathbf{u}^0$ such that $P^\varepsilon \to p^0$ strongly in $L^2_{\text{loc}}(\Omega)/\mathbb{R}$ and $\mathbf{U}^\varepsilon/\varepsilon^2 \to \mathbf{u}^0$ weakly in $L^2(\Omega)$. Finally, the homogenized pressure p^0 is shown to be a solution to the Darcy problem

$$\begin{aligned} \nabla \cdot a^0(\mathbf{f} - \nabla p^0) &= 0 \quad \text{in } \Omega, \\ a^0(\mathbf{f} - \nabla p^0) \cdot \mathbf{n} &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{2}$$

where the effective permeability a^0 is related to the porous structure of Ω_ε as is presented below. Moreover, we have $\mathbf{u}^0 = a^0(\mathbf{f} - \nabla p^0)$.

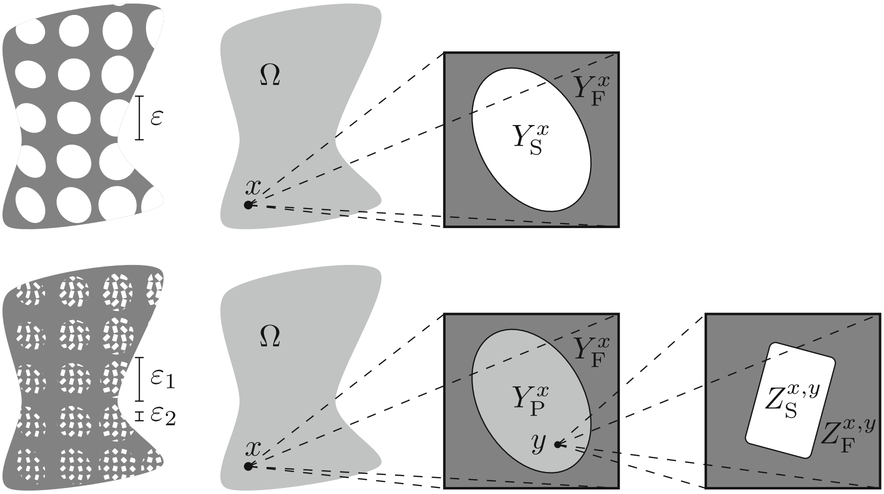


Fig. 1 The construction of Ω_ε (*top*) and $\Omega_{\varepsilon_1,\varepsilon_2}$ (*bottom*)

In the next two sections we describe the two- and three-scale porous media that are illustrated in Fig. 1.

2.1 Two-Scale Porous Media

We recall the definition of periodic and non-periodic two-scale porous media $\Omega_\varepsilon \subset \Omega$. Denote by Y the d-dimensional unit cube $(-1/2, 1/2)^d$, let $Y_S \subset \overline{Y}$, and set $Y_F = Y \backslash Y_S$. Here and subsequently the subscripts F and S stand for the fluid and solid part, respectively. We define a two-scale periodic porous medium in Ω by $\Omega_\varepsilon = \Omega \backslash \cup_{k \in \mathbb{Z}^d} \varepsilon(k + Y_S)$. Homogenization theory requires additional assumptions on Y_S and Y_F, but they are not too restrictive. We assume that Y_S is closed in $\overline{Y}$, both Y_S and Y_F have positive measure. Moreover, the sets Y_F and $\mathbb{R}^d \backslash \cup_{k \in \mathbb{Z}^d} (k + Y_S)$ are connected, have locally Lipschitz boundaries, and are locally located on one side of their boundaries.

We define non-periodic porous media by allowing for a deformation of the reference pore geometry. Consider a continuous map $\varphi : \mathbb{R}^d \times \overline{Y} \to \overline{Y}$ such that for every $x \in \mathbb{R}^d$ the function $\varphi(x, \cdot) : \overline{Y} \to \overline{Y}$ is a homeomorphism with $\varphi(x, \cdot), \varphi(x, \cdot)^{-1} \in W^{1,\infty}(Y)$. For any $x \in \Omega$ we define the local porous geometry as $Y_S^x = \varphi(x, Y_S)$ and $Y_F^x = Y \backslash Y_S^x$. We define a non-periodic two-scale porous medium by

$$\Omega_\varepsilon = \Omega \backslash \cup_{k \in \mathbb{Z}^d} \varepsilon(k + Y_S^{\varepsilon k}).$$

In the two-scale setting the homogenization theory relates the local porous geometry $(Y_{\mathrm{F}}^{x}, Y_{\mathrm{S}}^{x})$ to the effective permeability as follows. For any point $x \in \Omega$ and $i \in \{1, \dots, d\}$ we solve the Stokes micro problem: find the velocity field $\mathbf{u}^{i,x}$ and pressure $p^{i,x}$ such that

$$\begin{aligned} -\Delta \mathbf{u}^{i,x} + \nabla p^{i,x} &= \mathbf{e}^{i} \quad \text{in } Y_{\mathrm{F}}^{x}, & \mathbf{u}^{i,x} &= 0 \quad \text{on } \partial Y_{\mathrm{S}}^{x}, \\ \operatorname{div} \mathbf{u}^{i,x} &= 0 \quad \text{in } Y_{\mathrm{F}}^{x}, & \mathbf{u}^{i,x} \text{ and } p^{i,x} & \quad \text{are } Y\text{-periodic}, \end{aligned} \tag{3}$$

where $\mathbf{e}^{i}$ is the i-th canonical basis vector in $\mathbb{R}^{d}$. We then define

$$a_{ij}^{0}(x) = \int_{Y_{\mathrm{F}}^{x}} \mathbf{e}^{i} \cdot \mathbf{u}^{j,x} \, \mathrm{d}y \qquad \forall i, j \in \{1, \dots, d\}. \tag{4}$$

An explicit expression for $a^{0}(x)$ is generally unknown and must therefore be computed numerically using (3) and (4).

2.2 Three-Scale Porous Media

We consider porous media with a characteristic geometry at two different scales ε_1 and ε_2, where $\varepsilon_1 \gg \varepsilon_2 > 0$. If we apply the two-scale framework with $\varepsilon = \varepsilon_1$, parts of the micro domains Y_{F}^{x} will contain a characteristic geometry at scale $\varepsilon_2/\varepsilon_1 \ll 1$. In other words, a part (or whole) of Y_{F}^{x} can be considered as a porous medium with pores at scale $\varepsilon_2/\varepsilon_1$. In this situation, a direct numerical approximation of the micro problems (3) can become very costly, if not impossible.

We now embark in defining a three-scale porous medium $\Omega_{\varepsilon_1,\varepsilon_2} \subset \Omega$. Let us start with the description of the meso scale. Let $Y_{\mathrm{P}} \subset \overline{Y}$ and $Y_{\mathrm{F}} = Y \backslash Y_{\mathrm{P}}$, where P stands for porous part. We call $(Y_{\mathrm{F}}, Y_{\mathrm{P}})$ the reference mesoscopic geometry. To provide a variation at the meso scale we consider a continuous map $\varphi_1 : \mathbb{R}^{d} \times \overline{Y} \to \overline{Y}$ with the same properties as the map φ defined for two-scale porous media. For any $x \in \Omega$ we define the local mesoscopic geometry as $Y_{\mathrm{P}}^{x} = \varphi_1(x, Y_{\mathrm{P}})$ and $Y_{\mathrm{F}}^{x} = Y \backslash Y_{\mathrm{P}}^{x}$.

The porous structure of Y_{P}^{x} is described by the micro scale. Consider a continuous map $\varphi_2 : \mathbb{R}^{d} \times \mathbb{R}^{d} \times \overline{Y} \to \overline{Y}$ such that for every $x, y \in \mathbb{R}^{d}$ the map $\varphi_2(x, y, \cdot) : \overline{Y} \to \overline{Y}$ is a homeomorphism such that $\varphi_2(x, y, \cdot), \varphi_2(x, y, \cdot)^{-1} \in W^{1,\infty}(Y)$. Since we often fix parameters x and y, we simplify the notation by denoting a pair of x and y simply as $s = (x, y)$. That is, we can write $\varphi_2(x, y, z) \equiv \varphi_2(s, z)$. Let $Z_{\mathrm{S}} \subset \overline{Y}$ and $Z_{\mathrm{F}} = Y \backslash Z_{\mathrm{S}}$ be the miroscopic reference porous geometry. For any $s \in \mathbb{R}^{d} \times \mathbb{R}^{d}$ we define the local microscopic geometry as $Z_{\mathrm{S}}^{s} = \varphi_2(s, Y_{\mathrm{S}})$ and $Z_{\mathrm{F}}^{s} = Y \backslash Z_{\mathrm{S}}^{s}$.

For any $x \in \Omega$ we have now two different ways to view the local porous structure at x. First, we have the local mesoscopic geometry $(Y_{\mathrm{F}}^{x}, Y_{\mathrm{P}}^{x})$. Second, we can use the micro structure to define a fine scale description

$$\tilde{Y}_{\mathrm{S}}^{x} = Y_{\mathrm{P}}^{x} \backslash \cup_{k \in \mathbb{Z}^{d}} (\varepsilon_2/\varepsilon_1)(k + Z_{\mathrm{S}}^{x, \varepsilon_2 k/\varepsilon_1}), \qquad \tilde{Y}_{\mathrm{F}}^{x} = Y \backslash \tilde{Y}_{\mathrm{S}}^{x}.$$

Notice that $Y_{\mathrm{F}}^x \cup Y_{\mathrm{P}}^x = \tilde{Y}_{\mathrm{F}}^x \cup \tilde{Y}_{\mathrm{S}}^x$ and $\tilde{Y}_{\mathrm{S}}^x \subset Y_{\mathrm{P}}^x$ hence $Y_{\mathrm{F}}^x \subset \tilde{Y}_{\mathrm{F}}^x$. We define a three-scale porous medium in Ω by

$$\Omega_{\varepsilon_1,\varepsilon_2} = \Omega \setminus \cup_{k\in\mathbb{Z}^d}\, \varepsilon_1(k + \tilde{Y}_{\mathrm{S}}^{\varepsilon_1 k}).$$

A fluid flow in $\Omega_{\varepsilon_1,\varepsilon_2}$ can be modeled by the Stokes equation as in (1). If we apply a two-scale numerical method to the three-scale medium, we will need to solve the Stokes micro problems (3) in the domains $\tilde{Y}_{\mathrm{F}}^x$, that is: find the velocity field $\tilde{\mathbf{u}}^{i,x}$ and pressure $\tilde{p}^{i,x}$ such that

$$\begin{aligned} -\Delta \tilde{\mathbf{u}}^{i,x} + \nabla \tilde{p}^{i,x} &= \mathbf{e}^i \quad \text{in } \tilde{Y}_{\mathrm{F}}^x, & \tilde{\mathbf{u}}^{i,x} &= 0 \quad \text{on } \partial \tilde{Y}_{\mathrm{S}}^x, \\ \operatorname{div} \tilde{\mathbf{u}}^{i,x} &= 0 \quad \text{in } \tilde{Y}_{\mathrm{F}}^x, & \tilde{\mathbf{u}}^{i,x} \text{ and } \tilde{p}^{i,x} &\quad \text{are } Y\text{-periodic.} \end{aligned} \tag{5}$$

As we mentioned, a direct numerical solution to (5) might be infeasible due to the complexity of $\tilde{Y}_{\mathrm{F}}^x$. We overcome this issue by an approximation to (5) using again a homogenization-based approach. As a first attempt, one can try applying the Stokes model in the fluid part Y_{F}^x and the Darcy model in the porous part Y_{P}^x. The permeability at any $y \in Y_{\mathrm{P}}^x$ can be upscaled from the micro geometry $Z_{\mathrm{F}}^{x,y}$. However, the Stokes and Darcy models would need to be coupled at the interface of Y_{F}^x and Y_{P}^x. Such couplings, for example the Beavers–Joseph interface conditions, are non-trivial due to different orders of the models. We prefer a different approach that avoids interface conditions completely by using the Stokes–Brinkman equation at the mesocopic level. We thus consider the following mesoscopic problem: for any $x \in \Omega$ and $i \in \{1, \dots, d\}$ find the velocity $\mathbf{u}^{i,x}$ and pressure $p^{i,x}$ such that

$$\begin{aligned} -\Delta \mathbf{u}^{i,x} + \nabla p^{i,x} + K^0 \mathbf{u}^{i,x} &= \mathbf{e}^i \quad \text{in } Y, & \mathbf{u}^{i,x}, p^{i,x} &\quad \text{are } Y\text{-periodic,} \\ \operatorname{div} \mathbf{u}^{i,x} &= 0 \quad \text{in } Y, \end{aligned} \tag{6}$$

where

$$K^0(x,y) = \begin{cases} (\varepsilon_1/\varepsilon_2)^2 b^0(x,y)^{-1} & \text{if } y \in Y_{\mathrm{P}}^x \\ 0 & \text{otherwise.} \end{cases}$$

and the microscopic permeability $b^0(x,y)$ is defined below in (9). We set

$$a_{ij}^0(x) = \int_Y \mathbf{e}^i \cdot \mathbf{u}^{j,x}\, \mathrm{d}y, \qquad \forall i,j \in \{1,\dots,d\}. \tag{7}$$

The micro permeability tensor $b^0 : \Omega \times Y \to \mathbb{R}^{d\times d}$ depends on the micro porous structure. For any $s = (x,y) \in \Omega \times Y$ we can compute $b^0(s) = b^0(x,y)$ by solving the Stokes micro problems

$$\begin{aligned} -\Delta \mathbf{u}^{i,s} + \nabla p^{i,s} &= \mathbf{e}^i \quad \text{in } Z^s_{\mathrm{F}}, & \mathbf{u}^{i,s} &= 0 \quad \text{on } \partial Z^s_{\mathrm{S}}, \\ \operatorname{div} \mathbf{u}^{i,s} &= 0 \quad \text{in } Z^s_{\mathrm{F}}, & \mathbf{u}^{i,s} \text{ and } p^{i,s} & \quad \text{are } Y\text{-periodic} \end{aligned} \tag{8}$$

for thevelocity $\mathbf{u}^{i,s}$ and pressure $p^{i,s}$, where $i \in \{1, \dots, d\}$ and define

$$b^0_{ij}(s) = b^0_{ij}(x, y) = \int_{Y^x_{\mathrm{F}}} \mathbf{e}^i \cdot \mathbf{u}^{j,s} \, \mathrm{d}y \qquad \forall i, j \in \{1, \dots, d\}. \tag{9}$$

We have seen a two-scale and a three-scale model problem. In the two-scale problem we use the macroscopic Darcy model (2) and the microscopic Stokes model (3) with the effective permeability (4). In the three-scale problem we use the macroscopic Darcy model (2), the mesoscopic Stokes-Brinkman model (6), and the microscopic Stokes model (8) with the effective permeabilities (7) and (9).

3 Numerical Multiscale Methods

We briefly describe here the numerical multiscale methods developed in [2] to solve the model problems from Sect. 2. We start with the macro scale discretization, which is the same for both methods. In Sect. 3.1 we outline the discretization of the micro and meso problems, which are collectively called *cell problems*. A common framework to work with all cell problems is presented in Sect. 3.2.

Let $\{\mathcal{T}_H\}$ be a family of conformal, shape-regular triangulations of Ω parametrized by the mesh size $H = \max_{K \in \mathcal{T}_H} \operatorname{diam}(K)$. We consider the macro finite element space $S^l(\mathcal{T}_H)$ of degree $l \in \mathbb{N}$ given by

$$S^l(\mathcal{T}_H) = \{q \in H^1(\Omega);\ q|_K \in \mathcal{P}^l(K),\ \forall K \in \mathcal{T}_H\},$$

where $\mathcal{P}^l(K)$ is the space of polynomials of total degree l in element K. For every $K \in \mathcal{T}_H$ we considera quadrature formula $(x_{K_j}, \omega_{K_j})_{j=1,\dots,J_{\mathrm{mac}}}$ with integration points $x_{K_j} \in K$ and positive weights ω_{K_j}. To achieve the optimal order of accuracy we assume that $\int_K q(x)\,\mathrm{d}x = \sum_{j=1}^{J_{\mathrm{mac}}} \omega_{K_j} q(x_{K_j})$ for any $q \in \mathcal{P}^{l'}(K)$, where $l' = \max(2l-2, l)$. A direct application of the FE method to (2) reads as follows: find $p^H \in S^l(\mathcal{T}_H)/\mathbb{R}$ such that

$$B_H(p^H, q^H) = L_H(q^H) \qquad \forall q^H \in S^l(\mathcal{T}_H)/\mathbb{R},$$

where the discrete macro bilinear form and right-hand side are given by

$$\begin{aligned} B_H(p^H, q^H) &= \sum_{K\in\mathscr{T}_H} \sum_{j=1}^{J_{\text{mac}}} \omega_{K_j} a^{h_1}(x_{K_j}) \nabla p^H(x_{K_j}) \cdot \nabla q^H(x_{K_j}), \\ L_H(q^H) &= \sum_{K\in\mathscr{T}_H} \sum_{j=1}^{J_{\text{mac}}} \omega_{K_j} a^{h_1}(x_{K_j}) \mathbf{f}(x_{K_j}) \cdot \nabla q^H(x_{K_j}). \end{aligned} \tag{10}$$

The tensor a^{h_1} that appears in (10) is a numerical approximation of a^0 from (7) if we are in the three-scale settings. We use the tensor a^h (a numerical approximation of (4)) if we are in a two-scale setting.

3.1 Cell Problems Transformation and Discretization

We recall that by cell problems we mean either

- the mesoscopic problem in the three-scale method (6), (7),
- the microscopic problem in the three-scale method (8), (9),
- or the microscopic problem in the two-scale method (3), (4).

The cell problems share many similarities. First, the unknowns are always velocity and pressure. Stable FE discretization for such problems are well-known and we will pick the Taylor–Hood finite element pairs. Second, the pressure is unique only up to an additive constant. Third, the velocity fields are always integrated to obtain an effective parameter for the coarser scale, see (7), (4), (9). To discretize any cell problem we proceed in several steps.

1. A weak formulation is obtained with the help of a Lagrange multiplier to normalize the pressure in order to obtain a unique solution in finite element spaces of periodic functions.
2. A change of variables is performed to map the physical sampling domain to the reference domain (such as Y_{F} or Z_{F}).
3. A Taylor–Hood FE pair is used to discretize the problem.
4. A quadrature formula is used if permeability data need to be upscaled from a finer scale (this applies to the meso scale problem, where an approximation to b^0 will be evaluated only at quadrature points in Y_{P}).
5. A discrete approximation of the permeability to be upscaled is defined.

We briefly discuss the method developed in [5] for the meso scale problem (6) in the three-scale method and refer reader to [4] for a detailed description of the micro problems. The weak formulation of (6) with a Lagrange multiplier to normalize the pressure reads as follows: for any $x \in \Omega$ and $i \in \{1, \ldots, d\}$ find a velocity field

$\mathbf{u}^{i,x} \in H^1_{\mathrm{per}}(Y)^d$, a pressure $p^{i,x} \in L^2(Y)$, and a Lagrange multiplier $\lambda^{i,x} \in \mathbb{R}$ such that

$$
\begin{aligned}
&\int_Y \Big(\sum_{j=1}^{d} \nabla \mathbf{u}_j^{i,x} \cdot \nabla \mathbf{v}_j - p^{i,x} \operatorname{div} \mathbf{v} \Big) \, \mathrm{d}y \\
&\qquad + \int_Y K^0(x,y) \mathbf{u}^{i,x} \cdot \mathbf{v} \, \mathrm{d}y = \int_Y \mathbf{e}^i \cdot \mathbf{v} \, \mathrm{d}y \qquad \forall \mathbf{v} \in H^1_{\mathrm{per}}(Y), \\
&\int_Y (-q \operatorname{div} \mathbf{u}^{i,x} + \lambda^{i,x} q) \, \mathrm{d}y = 0 \qquad \forall q \in L^2(Y), \\
&\int_Y \kappa p^{i,x} \, \mathrm{d}y = 0 \qquad \forall \kappa \in \mathbb{R},
\end{aligned}
\tag{11}
$$

where the space $H^1_{\mathrm{per}}(Y)$ consists of Y-periodic functions from $H^1(Y)$. We map the problem (11) into the reference meso structure $(Y_{\mathrm{F}}, Y_{\mathrm{P}})$ by applying the change of variables $y_{\mathrm{old}} = \varphi_1(x, y_{\mathrm{new}})$. Next, we sum the three equations into one to obtain a compact form that acts in $X_{\mathrm{mes}} = H^1_{\mathrm{per}}(Y) \times L^2(Y) \times \mathbb{R}$. The resulting problem, which is symmetric and non-coercive, and the output of interest a^0 (see (7)) are given by: find $\mathbf{U}^{i,x} \in X_{\mathrm{mes}}$ such that

$$
A_{\mathrm{mes}}(\mathbf{U}^{i,x}, \mathbf{V}; x) = G^i_{\mathrm{mes}}(\mathbf{V}; x) \qquad \forall \mathbf{V} \in X_{\mathrm{mes}}, \tag{12}
$$

$$
a^0_{ij}(x) = G^i_{\mathrm{mes}}(\mathbf{U}^{j,x}; x) \qquad \forall i,j \in \{1, \ldots, d\}. \tag{13}
$$

where the bilinear form $A_{\mathrm{mes}}(\cdot, \cdot; x) : X_{\mathrm{mes}} \times X_{\mathrm{mes}} \to \mathbb{R}$ and the right-hand side $G^i_{\mathrm{mes}}(\cdot; x) : X_{\mathrm{mes}} \to \mathbb{R}$ contain integral terms with coefficients that depend on the Jacobian $\nabla_y \varphi_1(x, y)$.

We now discretize the problem (12). Let $\mathscr{T}_{h_1}$ be a conformal, shape-regular triangulation of Y, where $h_1 = \max_{K \in \mathscr{T}_{h_1}} \operatorname{diam}(K)$. We assume that for every $K \in \mathscr{T}_{h_1}$ we have either $K \subset Y_{\mathrm{F}}$ or $K \subset Y_{\mathrm{P}}$. Let $k \in \mathbb{N}$ and define the Taylor-Hood $\mathbb{P}^{k+1}/\mathbb{P}^k$ FE spaces given by

$$
\begin{aligned}
V^{h_1}_{\mathrm{mes}} &= \{ \mathbf{v} \in S^{k+1}(\mathscr{T}_{h_1})^d; \quad \mathbf{v} \text{ is } Y\text{-periodic} \}, \\
P^{h_1}_{\mathrm{mes}} &= \{ q \in S^{k}(\mathscr{T}_{h_1}); \qquad q \text{ is } Y\text{-periodic} \}.
\end{aligned}
$$

Let $X^{h_1}_{\mathrm{mes}} = V^{h_1}_{\mathrm{mes}} \times P^{h_1}_{\mathrm{mes}} \times \mathbb{R} \subset X_{\mathrm{mes}}$. For every $K \in \mathscr{T}_{h_1}$ we consider a quadrature formula $(y_{K_j}, \omega_{K_j})_{j=1,\ldots,J_{\mathrm{mes}}}$ with integration points $y_{K_j} \in K$ and positive weights ω_{K_j}. An optimal order of accuracy is achieved if $\int_K q(y) \, \mathrm{d}y = \sum_{j=1}^{J_{\mathrm{mes}}} \omega_{K_j} q(y_{K_j})$ for any $q \in \mathscr{P}^{2(k+1)}(K)$. A discretization of (12) then reads: For any $x \in \Omega$ and $i \in$

$\{1,\dots,d\}$ find $\mathbf{U}_{h_1}^{i,x} \in X_{\mathrm{mes}}^{h_1}$ such that

$$\begin{aligned} A_{\mathrm{mes}}^{h_1}(\mathbf{U}_{h_1}^{i,x}, \mathbf{V}; x) &= G_{\mathrm{mes}}^{i}(\mathbf{V}; x) && \forall \mathbf{V} \in X_{\mathrm{mes}}^{h_1}, \\ a_{ij}^{h_1}(x) &= G_{\mathrm{mes}}^{i}(\mathbf{U}_{h_1}^{j,x}; x) && \forall i,j \in \{1,\dots,d\}, \end{aligned} \tag{14}$$

where

$$\begin{aligned} A_{\mathrm{mes}}^{h_1}(\mathbf{U}, \mathbf{V}; x) &= A_{\mathrm{mes}}^{h_1}((\mathbf{u}, p, \lambda), (\mathbf{v}, q, \kappa); x) \\ &= \sum_{K \in \mathscr{T}_{h_1} \cap Y_{\mathrm{P}}} \sum_{j=1}^{J_{\mathrm{mes}}} \omega_{K_j} \frac{{\varepsilon_1}^2}{{\varepsilon_2}^2} (b^{h_2}(x, \varphi_1(x, y_{K_j})))^{-1} \mathbf{u}(y_{K_j}) \cdot \mathbf{v}(y_{K_j}) \, \mathrm{d}y \\ &\quad + \int_Y \sum_{i,j=1}^{d} \Big(\rho_{ij}(x,y) \frac{\partial \mathbf{u}}{\partial y_i} \cdot \frac{\partial \mathbf{v}}{\partial y_j} - \sigma_{ij}(x,y) \Big(\frac{\partial \mathbf{v}_i}{\partial y_j} q + \frac{\partial \mathbf{u}_i}{\partial y_j} p \Big) \Big) \, \mathrm{d}y \\ &\quad + \int_Y \tau(x,y)(\lambda q + \kappa q) \, \mathrm{d}y, \\ G_{\mathrm{mes}}^{i}(\mathbf{V}; x) &= G_{\mathrm{mes}}^{i}((\mathbf{v}, q, \kappa); x) = \int_Y \tau(x,y) \mathbf{e}^i \cdot \mathbf{v} \, \mathrm{d}y, \end{aligned} \tag{15}$$

where we denote the Jacobian $J = J(x,y) = \nabla_y \varphi_1(x,y)$ and define

$$\begin{aligned} \rho(x,y) &= \det(J)(J^\top J)^{-1}, \quad \sigma(x,y) = \det(J) J^{-\top}, \\ \tau(x,y) &= \det(J). \end{aligned} \tag{16}$$

In (15) we denoted by b^{h_2} the numerical approximation of the micro permeability b^0 defined in (9). While the formulation (15) can seem complicated, it suffices to keep in mind the compact formulation (14).

We can apply the same approach to all the cell problems. The micro problems need to be mapped to their respective micro domains (Y_{F} in the two-scale method and Z_{F} in the three-scale method). For the micro problems a quadrature formula is not required as there is not a finer scale than the micro scale. To summarize our numerical procedure we sketch both numerical multiscale methods in a diagram in Fig. 2.

3.2 General form of a Cell Problem

The various cell problems in our numerical models can be written in the following abstract form. Let $\mathscr{D}$ be parametric space of dimension at most $2d$ and X be a finite-dimensional Hilbert space. We are given a symmetric parametric bilinear form A :

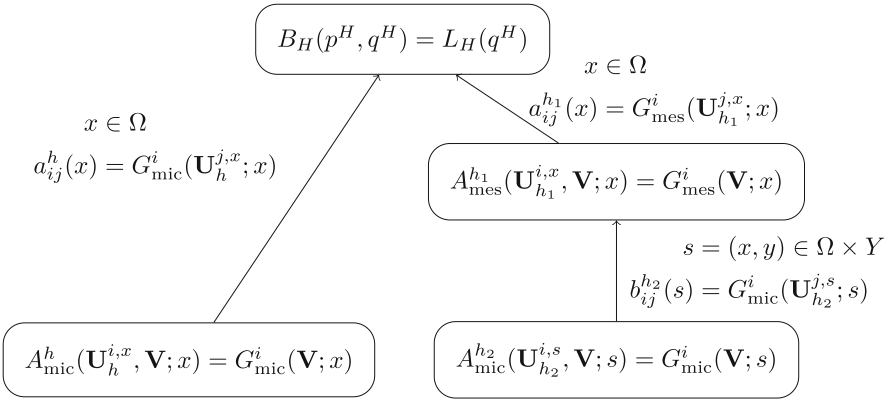

$B_H(p^H, q^H) = L_H(q^H)$
$x \in \Omega$
$a_{ij}^{h_1}(x) = G^i_{\text{mes}}(\mathbf{U}_{h_1}^{j,x}; x)$
$x \in \Omega$
$a_{ij}^{h}(x) = G^i_{\text{mic}}(\mathbf{U}_{h}^{j,x}; x)$
$A_{\text{mes}}^{h_1}(\mathbf{U}_{h_1}^{i,x}, \mathbf{V}; x) = G^i_{\text{mes}}(\mathbf{V}; x)$
$s = (x,y) \in \Omega \times Y$
$b_{ij}^{h_2}(s) = G^i_{\text{mic}}(\mathbf{U}_{h_2}^{j,s}; s)$
$A_{\text{mic}}^{h}(\mathbf{U}_{h}^{i,x}, \mathbf{V}; x) = G^i_{\text{mic}}(\mathbf{V}; x)$
$A_{\text{mic}}^{h_2}(\mathbf{U}_{h_2}^{i,s}, \mathbf{V}; s) = G^i_{\text{mic}}(\mathbf{V}; s)$

Fig. 2 A diagram of the two-scale method (*left branch*) and the three-scale method (*right branch*). Vertical direction: the Darcy macro scale (*top*), the Stokes-Brinkman meso scale (*middle*) and the Stokes micro scale (*bottom*)

$X \times X \times \mathscr{D} \to \mathbb{R}$ and parametric linear forms $G^i : X \times \mathscr{D} \to \mathbb{R}$ for $i \in \{1, \ldots, d\}$ with the inf-sup stability property

$$\inf_{\mathbf{U} \in X} \sup_{\mathbf{V} \in X} \frac{A(\mathbf{U}, \mathbf{V}; \mu)}{\|\mathbf{U}\|_X \|\mathbf{V}\|_X} \geq \beta(\mu) > 0 \qquad \forall \mu \in \mathscr{D}.$$

We are then interested in the evaluation of the output of interest $c : \mathscr{D} \to \mathbb{R}^{d \times d}$ that is defined via the following variational problems: for any $\mu \in \mathscr{D}$ and $i \in \{1, \ldots, d\}$ find $\mathbf{U}^{i,\mu} \in X$ such that

$$A(\mathbf{U}^{i,\mu}, \mathbf{V}; \mu) = G^i(\mathbf{V}; \mu), \qquad \forall \mathbf{V} \in X \tag{17}$$

$$c_{ij}(\mu) = G^i(\mathbf{U}^{j,\mu}) \qquad \forall i,j \in \{1, \ldots, d\}. \tag{18}$$

We see from Fig. 2 that all cell problems can be written in the form (17), (18).

4 Model-Order Reduction

Both the two and the three-scale methods presented in the previous section rely on the solution of a large number of cell problems of type (17) with different parameters and the construction of an upscaled permeability (18) to be used at a coarser scale. The effective permeability depends on a parameter in $\mathscr{D} = \Omega$ or $\mathscr{D} = \Omega \times Y$ of dimension at most $2d$, where d is the physical spatial dimension $d = 2, 3$. The repeated evaluation of the permeability for different values in $\mathscr{D}$ is a costly procedure as each evaluation relies on a PDE solve. Model order reduction can be used in this situation to build a low dimensional approximation of the solution manifold $\{\mathbf{U}^{i,\mu};\ \mu \in \mathscr{D}\}$. In our approach, we use the *reduced basis* (RB) method to construct such a low dimensional approximation space. The Petrov–Galerkin RB method [3] has been successfully applied to the two-scale problem [4] and to the three-scale problem in [5]. In Sect. 4.1 we present an abstract version ofthe RB methodology and apply it to the micro scale in Sect. 4.2 and to the meso scale in Sect. 4.3.

4.1 Petrov–Galerkin RB Method

For any $i \in \{1, \dots, d\}$ we construct a linear subspace $X_i \subset X$ that is spanned by a small number of solutions to (17). We then project (17) to the solution space X_i and a parameter-dependent test space $Y_i^\mu = T(X_i; \mu)$, where $T : X \times \mathscr{D} \to X$, called the supremizer operator, is defined below. The RB approximation of (17), (18) then reads: find $\mathbf{U}_{\mathrm{RB}}^{i,\mu} \in X_i$ such that

$$A(\mathbf{U}_{\mathrm{RB}}^{i,\mu}, \mathbf{V}; \mu) = G^i(\mathbf{V}; \mu) \qquad \forall \mathbf{V} \in Y_i^\mu. \tag{19}$$

We define a RB approximation of $c(\mu)$ with quadratic accuracy (see [15]) by

$$c_{ij}^{\mathrm{RB}}(\mu) = G^i(\mathbf{U}_{\mathrm{RB}}^{j,\mu}; \mu) + G^j(\mathbf{U}_{\mathrm{RB}}^{i,\mu}; \mu) - A(\mathbf{U}_{\mathrm{RB}}^{j,\mu}, \mathbf{U}_{\mathrm{RB}}^{i,\mu}; \mu). \tag{20}$$

For any $\mu \in \mathscr{D}$ and $\mathbf{U} \in X$ we define $T(\mathbf{U}; \mu) \in X$ as the unique element of X such that $(T(\mathbf{U}; \mu), \mathbf{V})_X = A(\mathbf{U}, \mathbf{V}; \mu)$ for every $\mathbf{V} \in X$. The supremizer operator $T(\mathbf{U}; \mu)$ is well-defined and linear in $\mathbf{U}$. Selecting Y_i^μ as the test space makes the method provably stable [3].

How do we construct a good solution space X_i? And how can we quickly evaluate (20) for any $\mu \in \mathscr{D}$? Answers to these questions rely on splitting the RB problem (19) and evaluating (20) at two different stages: an offline and an online stage.

- The *offline* stage is run only once and it is used to construct the RB space X_i and precompute necessary values for the online stage.
- The *online* stage can be run after the offline stage repeatedly and it provides a cheap and accurate approximation $c^{\mathrm{RB}}(\mu)$ for any $\mu \in \mathscr{D}$.

The RB space X_i is defined as the span of solutions $\mathbf{U}^{i,\mu}$ to (17) for a carefully selected small set of parameters $S^i \subset \mathscr{D}$, where $N_i \in \mathbb{N}$. Let us denote $(\mathbf{U}^{i,1}, \mathbf{U}^{i,2}, \dots, \mathbf{U}^{i,N_i})$ the result of applying the Gram–Schmidt orthogonalization procedure on these solutions. We thus have

$$X_i = \operatorname{span}\{\mathbf{U}^{i,1}, \mathbf{U}^{i,2}, \dots, \mathbf{U}^{i,N_i}\}.$$

The set S^i is constructed in the offline stage for every $i \in \{1, \ldots, d\}$ using a greedy algorithm. Given any S_i (even empty) and a corresponding space X_i, we can show that $\|\mathbf{U}^{i,\mu} - \mathbf{U}^{i,\mu}_{\mathrm{RB}}\|_X \leq \Delta^{\mathrm{E}}_i(\mu)$ for every parameter $\mu \in \mathscr{D}$, where the accurate a posteriori error estimator $\Delta^{\mathrm{E}}_i(\mu)$ can be evaluated cheaply for any $\mu \in \mathscr{D}$ (see [4] for details).

Algorithm: Greedy RB Construction. Select a training set $\varXi \subset \mathscr{D}$ and a tolerance $\varepsilon_{\mathrm{tol}} > 0$. For each $i \in \{1, \ldots, d\}$ we start with $S^i = \emptyset$ and repeat:

1. Find $\hat{\mu} \in \varXi$ for which the value $\Delta^{\mathrm{E}}_i(\hat{\mu})$ is the largest.
2. If $\Delta^{\mathrm{E}}_i(\hat{\mu}) < \varepsilon_{\mathrm{tol}}$, we stop. Else, we add $\hat{\mu}$ to S^i, update the space X_i, and continue with step 1.

The offline-online splitting requires an additional assumption: existence of an affine decomposition of A and G^i. Indeed, we assume that there exist $Q_A, Q_G \ll \dim(X)$ and

- symmetric bilinear forms $A^q(\cdot,\cdot) : X \times X \to \mathbb{R}$ for $q \in \{1, \ldots, Q_A\}$,
- linear forms $G^{iq}(\cdot) : X \to \mathbb{R}$ for $q \in \{1, \ldots, Q_G\}$ and $i \in \{1, \ldots, d\}$,
- vector fields $\varTheta^A : \mathscr{D} \to \mathbb{R}^{Q_A}$ and $\varTheta^G : \mathscr{D} \to \mathbb{R}^{Q_G}$,

such that for any $\mathbf{U}, \mathbf{V} \in X$, parameter $\mu \in \mathscr{D}$, and $i \in \{1, \ldots, d\}$ we have

$$\begin{aligned} A(\mathbf{U}, \mathbf{V}; \mu) &= \sum\nolimits_{q=1}^{Q_A} \varTheta^A_q(\mu) A^q(\mathbf{U}, \mathbf{V}), \\ G^i(\mathbf{V}; \mu) &= \sum\nolimits_{q=1}^{Q_G} \varTheta^G_q(\mu) G^{iq}(\mathbf{V}). \end{aligned} \tag{21}$$

One can then apply an affine decomposition (21) in the system (19) writing the RB solution as a linear combination $\mathbf{U}^{i,\mu}_{\mathrm{RB}} \in X_i$ in the form $\mathbf{U}^{i,\mu}_{\mathrm{RB}} = \sum_{n=1}^{N_i} \alpha^{i,\mu}_n \mathbf{U}^{i,n}$, where $\alpha^{i,\mu} = (\alpha^{i,\mu}_1, \ldots, \alpha^{i,\mu}_{N_i})^T \in \mathbb{R}^{N_i}$ is a vector of unknowns. This transformation yields a dense linear system of low dimension. This linear system can be assembled in the online stage in a time cost independent of $\dim(X)$ and the computation of solution $\alpha^{i,\mu}$ is usually very fast. Thus, we can obtain $\alpha^{i,\mu}$ without reconstructing the complete RB solution $\mathbf{U}^{i,\mu}_{\mathrm{RB}}$ and use this information in (20) to compute the output of interest $c^{\mathrm{RB}}(\mu)$, again with a time cost independent of $\dim(X)$.

4.2 RB Method at the Micro Scale

Micro problems in the two-scale and three-scale numerical methods are almost equivalent with the main difference being the parametric space. We have $\mathscr{D} = \varOmega$ in the two-scale model and $\mathscr{D} = \varOmega \times Y$ in the three-scale model. For simplicity of notation we consider just one of them, the three-scale model. Hence, we have a microscopic mesh size h_2, a microscopic reference mesh $\mathscr{T}_{h_2}$, a Hilbert space $X^{h_2}_{\mathrm{mic}}$

and for any parameter $s = (x, y) \in \Omega \times Y$ we have

$$
\begin{aligned}
A_{\text{mic}}^{h_2}(\mathbf{U}, \mathbf{V}; s) &= A_{\text{mic}}^{h_2}((\mathbf{u}, p, \lambda), (\mathbf{v}, q, \kappa); s) \\
&= \int_{Z_\text{F}} \sum_{i,j=1}^{d} \Big(\rho_{ij}(s, z) \frac{\partial \mathbf{u}}{\partial y_i} \cdot \frac{\partial \mathbf{v}}{\partial y_j} - \sigma_{ij}(s, z) \Big(\frac{\partial \mathbf{v}_i}{\partial y_j} q + \frac{\partial \mathbf{u}_i}{\partial y_j} p \Big) \Big) \, \mathrm{d}z \\
&\quad + \int_{Z_\text{F}} \tau(s, z)(\lambda q + \kappa q) \, \mathrm{d}z, \\
G_{\text{mic}}^{i}(\mathbf{V}; s) &= G_{\text{mic}}^{i}((\mathbf{v}, q, \kappa); s) = \int_{Z_\text{F}} \tau(s, z) \mathbf{e}^i \cdot \mathbf{v} \, \mathrm{d}z,
\end{aligned}
\tag{22}
$$

where we denote the Jacobian $J = J(s, z) = \nabla_z \varphi_2(s, z)$ and define the coefficients $\rho(s, z)$, $\tau(s, z)$, and $\sigma(s, z)$ exactly as in (16). To successfully apply the RB method, we need to construct an affine decomposition (21) of the forms $A_{\text{mic}}^{h_2}$ and G_{mic}^{i}. The main obstacle in doing so are the coefficients ρ_{ij}, σ_{ij}, and τ. If we could express them in the following affine form

$$
a_1(s) b_1(z) + \cdots + a_n(s) b_n(z) \tag{23}
$$

we could factor the s-dependent terms outside the integrals and an affine decomposition will be obtained. Decompositions of type (23) are not possible for arbitrary maps φ_2. However, if we assume that φ_2 is piecewise (in z) affine, then the Jacobian J will be piecewise constant, which yields a simple decomposition of type (23). Assuming that φ_2 is piecewise affine is a common practice in RB methodology for varying geometries. In case that this assumption is not valid, we can still rely on the empirical interpolation method (see Sect. 4.3) to obtain (23) at least approximately.

4.3 RB Method at the Meso Scale

The micro scale forms (22) and the meso scale forms (15) are very similar. They have the same terms containing ρ, τ, and σ that we dealt with in the previous section. Hence, it suffices to assume that φ_1 is piecewise affine (in y) and all but one term in (15) inherit an affine decomposition of the type (23). The only problematic term in the meso problem (15) is the term containing $b^{h_1}(x, \varphi_1(x, y))^{-1}$. Following the finding of [5] we apply the *empirical interpolation method* (EIM) [9] to obtain a decomposition

$$b^{h_1}(x, \varphi_1(x,y))^{-1} \approx a_1(x)b_1(y) + \cdots + a_n(x)b_n(y), \tag{24}$$

where the number of terms n controls the precision of the approximation. The EIM consists again of two stages: an offline stage and an online stage. The offline stage is a greedy algorithm that runs only once. The online stage allows a fast computation of the coefficients $a_1(x), a_2(x), \ldots, a_n(x)$ for any given $x \in \Omega$ by evaluating the left hand side of (24) for n selected values of y. To achieve the best performance in the three-scale method, one should combine the RB at meso and micro scale, which means that in (15) and in (24) we the tensor use b^{RB} instead of b^{h_2}.

5 Numerical Experiments

We illustrate the presented techniques with a two-scale numerical experiment. The code is implemented in Matlab and uses Matlab's `mldivide` to solve dense and sparse linear systems. We use $\mathbb{P}^2/\mathbb{P}^1$ Taylor–Hood FE on the micro scale and $\mathbb{P}^1$ FE on the macro scale.

Let the macroscopic domain Ω and the initial macroscopic mesh $\mathcal{T}_H$ be as depicted in Fig. 3 (left). We assume that the straight edges on the top and bottom of Ω are connected (periodic boundary conditions) and that the force field is constant with $\mathbf{f} \equiv (0, -1)$. The reference microscopic domain is depicted in Fig. 4. The domain Y_{F} contains four holes that represent solid obstacles. The domain deformation function φ can rotate the four obstacles around and uniformly scale their size and position. To illustrate the range of micro geometries, two examples of the deformed micro domains Y_{F}^x are provided in Fig. 4 . Moreover, we show how Y_{F} can be divided into nine parts such that φ is affine in each of them.

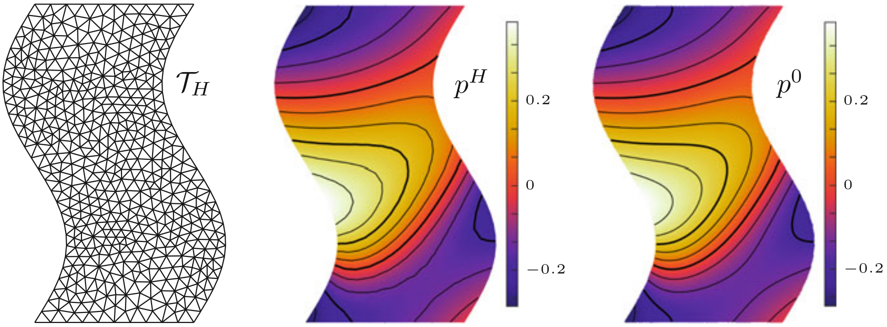


Fig. 3 Macroscopic mesh $\mathcal{T}_H$ (*left*), solution p^H to the reduced basis two-scale numerical method with 30 basis functions (*middle*), and an accurate approximation of p^0 (*right*)

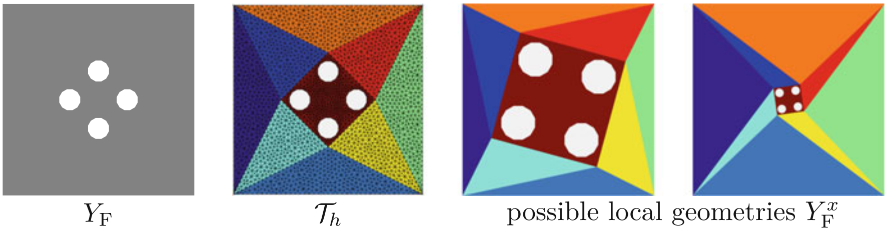


Fig. 4 From *left to right*: micro reference geometry Y_F, the microscopic mesh $\mathscr{T}_h$ and division of Y_F to nine subdomains, and two examples of a local porous geometry Y_F^x

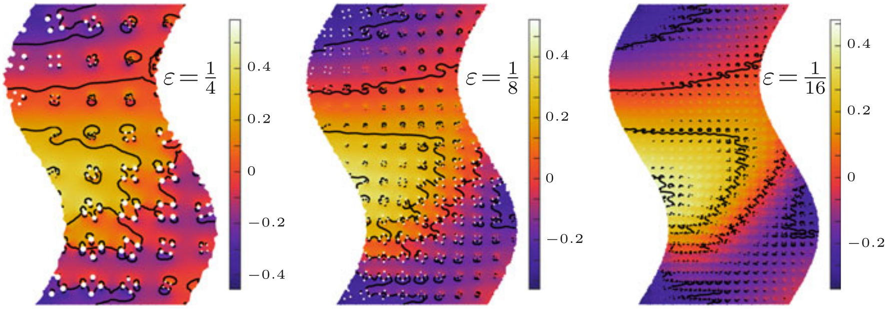


Fig. 5 Numerical approximations to the solutions p^ε to the fine scale problem (1) for $\varepsilon = 1/4, 1/8, 1/16$

In Fig. 5 we show the global variation of the porous structure for some (relatively large) values of ε and solutions to the fine scale problem (1). In the two-scale model we used reduced basis at the micro scale. Setting the tolerance of the greedy algorithm to $\varepsilon_{\text{tol}} = 0.01$ we obtained the reduced basis of size $N_1 = N_2 = 40$. The solution p^H is depicted in Fig. 3 along with a very accurate numerical reconstruction of p^0. The numerical solution p^H is in agreement with the fine scale solutions as can be seen in Fig. 5.

References

1. A. Abdulle, Y. Bai, Adaptive reduced basis finite element heterogeneous multiscale method. Comput. Methods Appl. Mech. Eng. **257**, 201–220 (2013)
2. A. Abdulle, O. Budáč, An adaptive finite element heterogeneous multiscale method for Stokes flow in porous media. Multiscale Model. Simul. **13**, 256–290 (2015)
3. A. Abdulle, O. Budáč, A Petrov–Galerkin reduced basis approximation of the Stokes equation in parameterized geometries. C. R. Math. Acad. Sci. Paris **353**(7), 641–645 (2015)
4. A. Abdulle, O. Budáč, A reduced basis finite element heterogeneous multiscale method for Stokes flow in porous media. Accept. Comp. Methods Appl. Mech. Eng. (2015)
5. A. Abdulle, O. Budáč, A three-scale offline-online numerical method for fluid flow in porous media (2016, preprint)

6. G. Allaire, Homogenization of the Stokes flow in a connected porous medium. Asymptot. Anal. **2**(3), 203–222 (1989)
7. G. Allaire, Homogenization of the Navier-Stokes equations in open sets perforated with tiny holes I. abstract framework, a volume distribution of holes. Arch. Ration. Mech. Anal. **113**(3), 209–259 (1991)
8. S. Alyaev, E. Keilegavlen, J.M. Nordbotten, Analysis of control volume heterogeneous multiscale methods for single phase flow in porous media. Multiscale Model. Simul. **12**(1), 335–363 (2014)
9. M. Barrault, Y. Maday, N.-C. Nguyen, A. T. Patera, An 'empirical interpolation method': application to efficient reduced-basis discretization of partial differential equations. C. R. Math. Acad. Sci. Paris **339**, 667–672 (2004)
10. D.L. Brown, Y. Efendiev, V.H. Hoang, An efficient hierarchical multiscale finite element method for Stokes equations in slowly varying media. Multiscale Model. Simul. **11**(1), 30–58 (2013)
11. D.L. Brown, P. Popov, Y. Efendiev, On homogenization of Stokes flow in slowly varying media with applications to fluid-structure interaction. GEM Int. J. Geomath. **2**(2), 281–305 (2011)
12. H. Darcy, Les fontaines publiques de la ville de Dijon: Exposition et application á suivre et des formules á employer dans les questions de duistribution deau (1856)
13. M. Griebel, M. Klitz, Homogenization and numerical simulation of flow in geometries with textile microstructures. Multiscale Model. Simul. **8**(4), 1439–1460 (2010)
14. V.H. Hoang, C. Schwab, High-dimensional finite elements for elliptic problems with multiple scales. Multiscale Model. Simul. **3**(1), 168–194 (2005)
15. N.A. Pierce, M.B. Giles, Adjoint recovery of superconvergent functionals from PDE approximations. SIAM Rev. **42**(2), 247–264 (2000)
16. E. Sánchez-Palencia, *Non-homogeneous Media and Vibration Theory*. Lecture Notes in Physics, vol. 127 (Springer, Berlin, 1980)
17. L. Tartar, Appendix, in *Incompressible Fluid Flow in a Porous Medium—Convergence of the Homogenization Process*. Lecture Notes in Physics, vol. 127, [16], 1979, pp. 368–377

Output Error Estimates in Reduced Basis Methods for Time-Harmonic Maxwell's Equations

Martin W. Hess and Peter Benner

Abstract The Reduced Basis Method (RBM) (Rozza et al., Archiv Comput Methods Eng 15:229–275, 2008) generates low-order models for efficient evaluation of parametrized PDEs in many-query and real-time contexts. It can be seen as a parametric model reduction method (Benner et al., SIAM Rev 57(4):483–531, 2015), where greedy selection is combined with a projection space composed of solution snapshots. The approximation quality is certified by using rigorous error estimators.

We apply the RBM to systems of Maxwell's equations arising from electrical circuits. Using microstrip models, the input-output behaviour of the interconnect structures is approximated for a certain frequency range. Typically, an output is given by a linear functional, but in the case of impedance parameters (also called Z-parameters), the output is quadratic. An expanded formulation is used to rewrite the system to compliant form, i.e., a form, where the input and output are identical. This enables fast convergence in the approximation error and thus very low reduced model sizes. A numerical example from the microwave regime shows the advantage of this approach.

1 Time-Harmonic Maxwell's Equations

We consider the time-harmonic Maxwell's equations in weak form over the discrete, high-dimensional function space X, discretized with Nédélec finite elements of first order [3, 4]

$$\mu^{-1}(\nabla \times E, \nabla \times v) + i\omega\sigma(E, v) - \omega^2\epsilon(E, v) = i\omega J, \quad \forall v \in X, \tag{1}$$

M.W. Hess (✉) • P. Benner
Max Planck Institute for Dynamics of Complex Technical Systems, Sandtorstr. 1,
39106 Magdeburg, Germany
e-mail: hessm@mpi-magdeburg.mpg.de; benner@mpi-magdeburg.mpg.de

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_33

subject to zero boundary conditions

$$E \times n = 0 \quad \text{on } \partial\Omega, \tag{2}$$

which is solved for the discretized electric field E in the computational domain Ω. A test function is denoted by v and $(\cdot,\cdot)$ is the complex $L^2(\Omega)$ scalar product. The source current density is denoted by J, the imaginary number i, the frequency ω and the material coefficients are the permeability μ, conductivity σ and permittivity ϵ.

Since the reduced basis method (RBM) makes use of the affine parameter dependence in ω, system matrices A_μ, A_σ and A_ϵ are assembled entry-wise using the finite element basis functions $\{\varphi_i, i = 1 \dots \mathcal{N}\}$, as

$$A_{\mu(i,j)} \equiv \mu^{-1}(\nabla \times \varphi_i, \nabla \times \varphi_j), \tag{3}$$

$$A_{\sigma(i,j)} \equiv \sigma(\varphi_i, \varphi_j), \tag{4}$$

$$A_{\epsilon(i,j)} \equiv \epsilon(\varphi_i, \varphi_j). \tag{5}$$

The complex system matrix is then composed as $A_\mu + i\omega A_\sigma - \omega^2 A_\epsilon$. Splitting the state vector into real and complex parts as $x_{real} + ix_{imag}$ and denoting the source term by b_J, leads to

$$(A_\mu + i\omega A_\sigma - \omega^2 A_\epsilon)(x_{real} + ix_{imag}) = i\omega b_J. \tag{6}$$

This allows a reformulation to a real symmetric system [5], as

$$\begin{bmatrix} \frac{1}{\omega}A_\mu - \omega A_\epsilon & -A_\sigma \\ -A_\sigma & -\frac{1}{\omega}A_\mu + \omega A_\epsilon \end{bmatrix} \begin{bmatrix} x_{real} \\ x_{imag} \end{bmatrix} = \begin{bmatrix} 0 \\ -b_J \end{bmatrix} \tag{7}$$

This system also permits an affine expansion in the frequency, namely

$$A(\omega) = \sum_{q=1}^{3} \Theta_a^q(\omega) A^q = \frac{1}{\omega} A^1 + A^2 + \omega A^3, \tag{8}$$

whereby the matrices in the affine form A^1, A^2 and A^3 are also real symmetric. This allows to use an implementation of the model reduction procedure in real arithmetic. However, a doubling of the system size comes from the transition to a real system (7), which is denoted by $A(\omega)x(\omega) = b$ subsequently.

The quantity $s(\omega) = |\ell(E)|$ for a linear functional $\ell(\cdot)$ acting on the complex electric field E serves as output,

$$s(\omega) = \sqrt{(\ell(\Re E))^2 + (\ell(\Im E))^2}. \tag{9}$$

Using the real form where ℓ_1 denotes the action of ℓ on the real part and ℓ_2 denotes the action of ℓ on the imaginary part, it follows

$$s(\omega) = \sqrt{\ell_1(x(\omega))^2 + \ell_2(x(\omega))^2}, \tag{10}$$

or in vector notation

$$s(\omega) = \sqrt{(\ell_1^T x)^2 + (\ell_2^T x)^2}. \tag{11}$$

Define the quadratic form $Q(\cdot,\cdot)$ by $\ell_1^T \ell_1 + \ell_2^T \ell_2$, it holds

$$s^2(\omega) = Q(x,x), \tag{12}$$

i.e., the squared output has a representation as a quadratic form.

2 Reduced Basis Method

The central idea of the reduced basis method (RBM) is that the parametric manifold $\mathscr{M} = \{E(\nu)|\nu \in \mathscr{D}\}$ can be well approximated by a linear space of snapshot solutions V_N of dimension N. A greedy-max sampling is employed to select the snapshot locations. A so-called offline-online decomposition enables the computational feasibility, in that a time-consuming offline phase generates V_N, while a fast online phase computes the output quantities using the reduced order model.

The real symmetric system (7) defines a set of parameter-dependent linear systems

$$A(\omega)x(\omega) = b, \tag{13}$$

with affine parameter dependence

$$A(\omega) = \sum_{q=1}^{Q_a} \Theta_a^q(\omega) A^q = \frac{1}{\omega} A^1 + A^2 + \omega A^3. \tag{14}$$

The affine parameter dependence is required for the offline-online decomposition. As this example considers only an expansion in the frequency, the affine form is readily established, see [6] on the treatment of geometric parameters for instance.

The greedy sampling uses an error estimator $\Delta_N(\omega)$, which estimates the error between the full order solution $E(\omega)$ and reduced order solution $E_N(\omega)$. It is assumed that the parameter domain is sampled in Ξ and an approximation tolerance ϵ is set.

```
Choose $\omega_1 \in \Xi$ arbitrarily
Solve (1) for $E(\omega_1)$
Set $S_1 = \{\omega_1\}$
Set $V_1 = [x(\omega_1)]$
Set $N = 1$
while $\max_{\omega \in \Xi} \Delta_N(\omega) \geq \epsilon$ do
    Set $\omega_{N+1} = \arg\max_{\omega \in \Xi} \Delta_N(\omega)$
    Solve (1) for $E(\omega_{N+1})$
    Set $S_{N+1} = S_N \cup \omega_{N+1}$
    Set $V_{N+1} = [V_N \quad x(\omega_{N+1})]$
    Orthonormalize the columns of $V_{N+1}$
    Set $N = N + 1$
end while
```

The projection onto the snapshot space V_N is carried out on the parameter independent parts, as

$$A_N^q = V_N^T A^q V_N, \tag{15}$$

$$b_N = V_N^T b. \tag{16}$$

This is a parameter-preserving model reduction, since the affine form is also present in the reduced system

$$\left(\sum_{q=1}^{Q_a} \Theta_a^q(\omega) A_N^q \right) x_N(\omega) = b_N. \tag{17}$$

Different choices for the error estimator $\Delta_N(\omega)$ are possible. The error estimator in the field is given by

$$\Delta_N(\omega) = \frac{\|r^{pr}(\cdot;\omega)\|_{X'}}{\beta_{LB}(\omega)}, \tag{18}$$

where $\|r^{pr}(\cdot;\omega)\|_{X'}$ denotes the dual norm of the residual of the primal problem (see [1] for the efficient computation) and $\beta_{LB}(\omega)$ is a lower bound to the inf-sup stability constant (see [7, 8] for approximations to this computationally expensive bound).

For linear outputs, a dual system is considered, which allows to define the output error estimator

$$\Delta_N^o(\omega) = \frac{\|r^{pr}(\cdot;\omega)\|_{X'} \|r^{du}(\cdot;\omega)\|_{X'}}{\beta_{LB}(\omega)}, \tag{19}$$

with the dual norm of the residual of the dual problem $\|r^{du}(\cdot;\omega)\|_{X'}$. The output error estimator computes estimates on the output error $|s(\omega) - s_N(\omega)|$, and generally leads to faster convergence in the output [1]. It is however not directly applicable to the quadratic output.

A special case is the compliant case, where input and output functional are identical, i.e., $\ell = f$, [1]. In this case, an output estimator is given by

$$\Delta_N^s(\omega) = \frac{\|r^{pr}(\cdot;\omega)\|_{X'}^2}{\beta_{LB}(\omega)}. \tag{20}$$

Using an expanded formulation as shown in [9] allows to transform the system with quadratic output into a compliant system. Consider

$$\mathscr{A}(\omega) = \begin{bmatrix} 2A(\omega) - Q & -Q \\ -Q & 2A(\omega) - Q \end{bmatrix} \quad \text{and} \quad \mathscr{B} = \begin{bmatrix} b \\ -b \end{bmatrix}. \tag{21}$$

Then, for the parametric problem $\mathscr{A}(\omega)\tilde{x}(\omega) = \mathscr{B}$, it holds $s^2(\omega) = \mathscr{B}\tilde{x}$. The disadvantage lies in a doubled system size, but due to the compliancy, a fast error decay is expected, which promises reduced models of lower order than when using a field estimator.

Many models from electromagnetics contain resonances, i.e., there are eigenfrequencies in the parameter domain where $A(\omega)$ is singular. This case corresponds to a zero stability constant $\beta(\omega) = 0$. In the model reduction context, the reduced system should preserve the eigenfrequencies or at least not introduce further eigenfrequencies. With the one-sided projection (15), the case that the stability constant of the reduced system $\beta_N(\omega) = 0$ while $\beta(\omega) > 0$ is often observed.

The relation that $\beta_N(\omega) \geq \beta(\omega)$ can be established by using a suitable two-sided projection. While the trial space is still set to the snapshot space $V_N = \{x(\omega_1), \ldots, x(\omega_N)\}$, a parameter-dependent test space W_N^ω is used. The test space is defined by applying the supremizing operators T^ω to each element of V_N, where $T^\omega = M^{-1}A(\omega)$, with M the inner product matrix of the space X, see also [8] for more details.

Summarizing, we consider the standard system

$$A(\omega)x(\omega) = b, \quad s^2(\omega) = \left(\ell_1^T x(\omega)\right)^2 + \left(\ell_2^T x(\omega)\right)^2 \tag{22}$$

and the expanded system

$$\mathscr{A}(\omega)\tilde{x}(\omega) = \mathscr{B}, \quad s^2(\omega) = \mathscr{B}^T\tilde{x}(\omega). \tag{23}$$

3 Numerical Results

A coplanar waveguide serves as numerical example. The model is shown in Fig. 1. The real symmetric form contains 2,024 degrees of freedom. This is a very coarse discretization for a 3D model and mainly serves to compare the different model

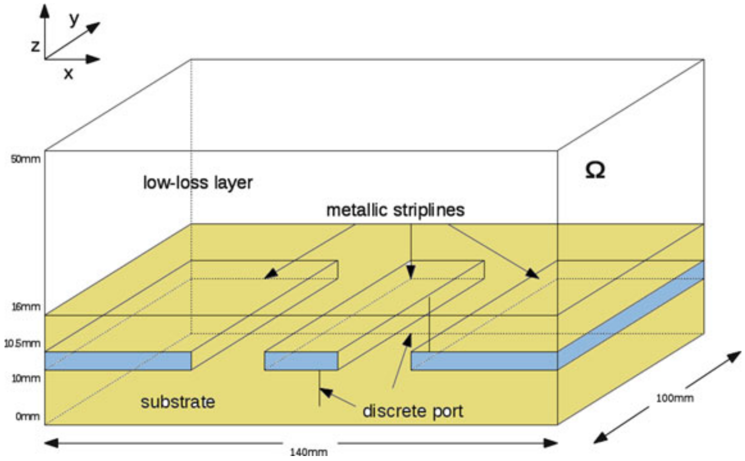


Fig. 1 Geometry of coplanar waveguide

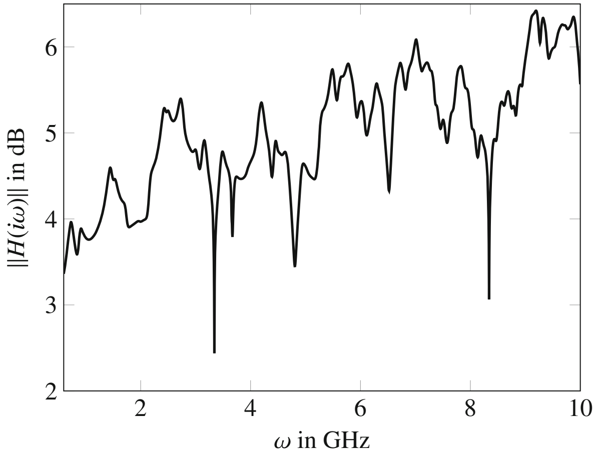


Fig. 2 Transfer function over frequency range [0.6, 10] GHz

reduction approaches. It can be expected that some physical features are not captured by this coarse discretization. The discrete ports shown in Fig. 1 serve as the input and output ports. In the upper part of the model ($z > 16$ mm), the relative permittivity is $\epsilon_r = 1.07$ and the conductivity is $\sigma = 0.01$ S/m. In the lower part ($z \leq 16$ mm), the relative permittivity is $\epsilon_r = 4.4$ and the conductivity is $\sigma = 0.02$ S/m. The relative permeability is one in the entire domain. The dimensions of the shielded box are 140 mm by 100 mm by 50 mm. The considered parameter is the frequency in [0.6, 10.0] GHz. The magnitude of the transfer function is shown in Fig. 2. It is the input-output mapping $\omega \mapsto s(\omega)$ and is denoted with $s(\omega) = \|H(i\omega)\|$ in dB in the figure.

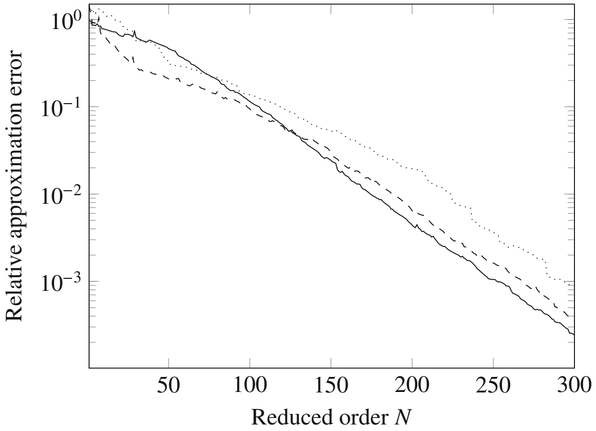


Fig. 3 Mean relative error over sampled grid. Field estimator (*dotted*), output estimator using expanded form (*solid*), heuristic optimum (*dashed*)

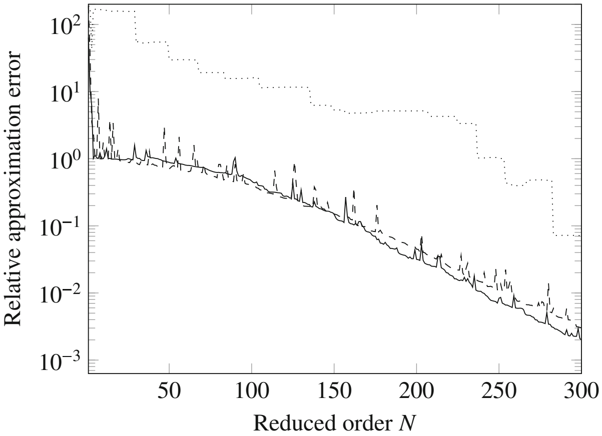


Fig. 4 Maximum relative error over sampled grid. Field estimator (*dotted*), output estimator using expanded form (*solid*), heuristic optimum (*dashed*)

The outcome of the reduced basis model reduction is shown in Figs. 3 and 4. The mean and maximum relative error in the output is plotted against the reduced basis size for three different cases. The greedy using the field estimator shows the slowest convergence but has the advantage to work with a system of standard size and is thus still advantageous when offline timing is important. The greedy using the output estimator and working with the expanded system shows fast convergence and is thus advantageous when the goal is a very small reduced order model. For comparison a 'heuristic optimum' or 'ideal greedy' is plotted, which uses the system of standard size with the actual errors in the output within the greedy. The fact that the expanded system gives comparable accuracy to the ‘ideal greedy’ strengthens the point, that the expanded system is beneficial when sufficient offline time is available.

Acknowledgements The model used for the numerical experiment is available through the MOR Wiki found under www.modelreduction.org. This work is supported by the collaborative project nanoCOPS, Nanoelectronic Coupled Problems Solutions, supported by the European Union in the FP7-ICT-2013-11 Program under Grant Agreement Number 619166.

References

1. G. Rozza, D.B.P. Huynh, A.T. Patera, Reduced basis approximation and a posteriori error estimation for affinely parametrized elliptic coercive partial differential equations. Archiv. Comput. Methods Eng. **15**, 229–275 (2008)
2. P. Benner, S. Gugercin, K. Willcox, A survey of projection-based model reduction methods for parametric dynamical systems. SIAM Rev. **57**(4), 483–531 (2015)
3. R. Hiptmair, Finite elements in computational electromagnetism. Acta Numer. **11**, 237–339 (2002)
4. S. Zaglmayr, High order finite element methods for electromagnetic field computation. PhD thesis, JKU Linz (2006)
5. M.W. Hess, P. Benner, Fast evaluation of time-harmonic Maxwell’s equations using the reduced basis method. IEEE Trans. Microw. Theory Tech. **61**(6), 2265–2274 (2013)
6. M.W. Hess, P. Benner, A reduced basis method for microwave semiconductor devices with geometric variations. COMPEL Int. J. Comput. Math. Electr. Electr. Eng. **33**(4), 1071–1081 (2014)
7. J.S. Hesthaven, G. Rozza, B. Stamm, *Certified Reduced Basis Methods for Parametrized Partial Differential Equations* (Springer, Cham, 2016)
8. M.W. Hess, S. Grundel, P. Benner, estimating the inf-sup constant in reduced basis methods for time-harmonic Maxwell’s equations. IEEE Trans. Microw. Theory Tech. **63**(11), 3549–3557 (2015)
9. S. Sen, Reduced basis approximation and a posteriori error estimation for non-coercive elliptic problems: application to acoustics. PhD thesis, Massachusetts Institute of Technology (2007)

Reduced Basis Exact Error Estimates with Wavelets

Mazen Ali and Karsten Urban

Abstract A (multi-)wavelet expansion is used to derive a rigorous bound for the (dual) norm Reduced Basis residual. We show theoretically and numerically that the error estimator is online efficient, reliable and rigorous. It allows to control the exact error (not only with respect to a "truth" discretization).

1 Introduction

The Reduced Basis Method (RBM) is a widely used mathematical framework for model reduction of parameterized partial differential equations (PPDEs). We refer to [5, 7] for recent books. One possible criticism of the RBM is that the reduction is based upon an a priorily fixed, so called "truth" discretization, which is assumed to be sufficiently fine in order to resolve the desired solution sufficiently well for all possible parameters. This means that an RB approximation can only be as good as the underlying truth. If this truth is not so "true", the RBM cannot be expected to yield a good approximation of the exact solution.

Thus, one would like to have an error estimator for the RB approximation with respect to the (unknown and not computable) exact solution of the PPDE. In this paper, we build upon a recent preprint [1], where we propose to use adaptive computations for the construction of the reduced model. In the present paper, we use such an adaptive method build upon (multi-)wavelets to construct an error estimator for the exact error. We show that this estimator is computable online efficient and gives sharp estimates (the latter statement is shown by numerical experiments).

In Sect. 2, we review the error estimates based upon the dual norm of the residual within the RBM, Sect. 3 is devoted to the wavelet-based error estimator and our numerical results are shown in Sect. 4.

M. Ali • K. Urban (✉)
Institute of Numerical Mathematics, Ulm University, Helmholtzstr. 20, D-89081 Ulm, Germany
e-mail: mazen.ali@uni-ulm.de; karsten.urban@uni-ulm.de

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_34

2 Reduced Basis Method (RBM) Error Estimation

In order to shorten notation, we consider parametric (for a parameter μ in a parameter set $\mathscr{D} \subset \mathbb{R}^P$) elliptic boundary value problems on $\mathscr{X} := H_0^1(\Omega)$, i.e., we look for $u(\mu) \in \mathscr{X}$ such that[1]

$$a(u(\mu), v; \mu) = \langle f(\mu), v \rangle \qquad \forall v \in \mathscr{X}, \tag{1}$$

where $f(\mu) \in \mathscr{X}' = H^{-1}(\Omega)$ is given and $\langle \cdot, \cdot \rangle$ denotes the duality pairing of $\mathscr{X}'$ and $\mathscr{X}$ with pivot space $L_2(\Omega)$. The bilinear form $a(\cdot, \cdot; \mu)$ is assumed to be symmetric, coercive and bounded with constants $\alpha(\mu)$ and $\gamma(\mu)$, respectively.

2.1 *Residual-Based Error Estimators*

Typically, RB error estimates are *residual-based*, where the *residual* $r(w; \mu) \in \mathscr{X}'$, $\mu \in \mathscr{D}$, is defined for given $w \in \mathscr{X}$ by

$$\langle r(w; \mu), v \rangle := \langle f(\mu), v \rangle - a(w, v; \mu), \qquad v \in \mathscr{X}. \tag{2}$$

The equivalence of error and residual is straightforward and well-known

$$\alpha(\mu)\, \|u(\mu) - w\|_{\mathscr{X}} \leq \|r(w; \mu)\|_{\mathscr{X}'} \leq \gamma(\mu)\, \|u(\mu) - w\|_{\mathscr{X}}, \quad w \in \mathscr{X}. \tag{3}$$

If an approximation space $X_N \subset \mathscr{X}$ of small dimension $N \in \mathbb{N}$ is constructed and a (Galerkin) approximation $u_N(\mu) \in X_N$ for a given parameter value $\mu \in \mathscr{D}$ has been computed, we set

$$R_N(\mu) := \|r(u_N(\mu); \mu)\|_{\mathscr{X}'} = \sup_{v \in \mathscr{X}} \frac{\langle r(u_N(\mu); \mu), v \rangle}{\|u_N(\mu)\|_{\mathscr{X}}}, \tag{4}$$

i.e., the *dual* norm of the residual. Usually, this dual norm is not computable, in particular since the supremum in (4) is taken over the *infinite*-dimensional space $\mathscr{X}$. Based upon $R_N(\mu)$, the error estimator for the *exact error* $u(\mu) - u_N(\mu)$ reads

$$\|u(\mu) - u_N(\mu)\|_{\mathscr{X}} \leq \frac{1}{\alpha(\mu)} R_N(\mu) =: \Delta_N(\mu). \tag{5}$$

Hence, one also needs the coercivity constant $\alpha(\mu)$, e.g. by the Successive Constraint Method (SCM), [6], which, however, is not the topic of this paper.

[1] We would like to stress that most what is said here can also be extended to Petrov-Galerkin inf-sup-stable problems with different trial and test spaces.

2.2 (Theoretical) Computing via Affine Decomposition and Riesz Representation

So far, any known computational procedure for the computation of $R_N(\mu)$ is based upon the assumption that the bilinear form $a(\cdot,\cdot;\mu)$ and the right-hand side allow for a separation of parameters and variables (often – a bit misleading – called an *affine decomposition* in the parameter), i.e.,

$$a(w,v;\mu) = \sum_{q=1}^{Q_a} \vartheta_q^a(\mu)\, a_q(w,v), \qquad f(\mu) = \sum_{q=1}^{Q_f} \vartheta_q^f(\mu) f_q, \tag{6}$$

where $\vartheta_q^a, \vartheta_q^f : \mathscr{D} \to \mathbb{R}$ and $a_q : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ as well as $f_q \in \mathscr{X}'$ are parameter-independent.[2] Then, it is straightforward to show that also the residual is affine in the parameter, i.e., with parameter-independent $r_q : \mathscr{X} \to \mathscr{X}'$, we have

$$r(w;\mu) = \sum_{q=1}^{Q_r} \vartheta_q^r(\mu)\, r_q(w), \qquad w \in \mathscr{X}. \tag{7}$$

Keeping in mind that $\mathscr{X}'$ is a Hilbert space with inner product $(\cdot,\cdot)_{\mathscr{X}'}$, one can try to compute $R_N(\mu)$ by using the expansion of the RB-solution in terms of the basis functions $\{\xi_1,\dots,\xi_N\}$ (to be specified later) of X_N, i.e.,

$$u_N(\mu) = \sum_{i=1}^{N} c_i(\mu)\, \xi_i, \qquad c_i(\mu) \in \mathbb{R}, \tag{8}$$

as follows[3]

$$\begin{aligned} R_N(\mu)^2 &= \|r(u_N(\mu);\mu)\|_{\mathscr{X}'}^2 = \big(r(u_N(\mu);\mu), r(u_N(\mu);\mu)\big)_{\mathscr{X}'} \\ &= \sum_{q,q'=1}^{Q_r} \sum_{i,i'=1}^{N} \vartheta_q^r(\mu)\, \vartheta_{q'}^r(\mu)\, c_i(\mu)\, c_{i'}(\mu)\, \big(r_q(\xi_i), r_{q'}(\xi_{i'})\big)_{\mathscr{X}'}. \end{aligned} \tag{9}$$

Obviously, the inner products

$$\boldsymbol{R}_{(q,i),(q',i')} := \big(r_q(\xi_i), r_{q'}(\xi_{i'})\big)_{\mathscr{X}'} \tag{10}$$

[2] If (6) does not hold, the *Empirical Interpolation Method (EIM)* determines an approximation with an upper interpolation bound [2], which, however, may not be accessible and which also reduces the sharpness of the error bound.

[3] Note, that (9) amounts to take the square root, which is not a problem in terms of efficiency, but it is an issue with respect to accuracy – the well-known so-called "square root effect".

are parameter-independent. If these values can be precomputed (in an offline phase), (9) amounts $(Q_r N)^2$ terms so that $R_N(\mu)$ can be computed with complexity $\mathcal{O}(N^2)$ – independent of $\mathcal{N}$, which is called *online efficient*.

A possible way to compute the terms in (10) is by determining the *Riesz representations* $\hat{r}_{q,i} \in \mathcal{X}$ of $r_q(\xi_i) \in \mathcal{X}'$, that are given by

$$\left(\hat{r}_{q,i}, v\right)_{\mathcal{X}} = \langle r_q(\xi_i), v \rangle \qquad \forall v \in \mathcal{X}, \tag{11}$$

where $(\cdot, \cdot)_{\mathcal{X}}$ denotes the inner product in $\mathcal{X}$. Once these Riesz representations are determined, one can compute (10) by $\boldsymbol{R}_{(q,i),(q',i')} = \left(\hat{r}_{q,i}, \hat{r}_{q',i'}\right)_{\mathcal{X}}$. Doing so, one avoids the dual inner product. Moreover, all computations for $\boldsymbol{R}_{(q,i),(q',i')}$ can be done offline. One only has to store $(Q_r N)^2$ numbers which are combined with the parameter-dependent terms in (9). However, this approach is only theoretically feasible since the computation of the Riesz representations in (11) would amount solving an *infinite-dimensional* problem.

2.3 The "Truth"

In a standard RBM, the way-out is through a common detailed discretization $\mathcal{X}^{\mathcal{N}} \subset \mathcal{X}$, where $\mathcal{N} \in \mathbb{N}$ is the dimension of the "truth" space, which is assumed to be rich enough to resolve the unknown $u(\mu)$ sufficiently accurate for all parameters $\mu \in \mathcal{D}$, i.e., the error $\|u(\mu) - u^{\mathcal{N}}(\mu)\|_{\mathcal{X}}$ is sufficiently small, where $u^{\mathcal{N}}(\mu) \in \mathcal{X}^{\mathcal{N}}$ is the corresponding truth approximation. This detailed space $\mathcal{X}^{\mathcal{N}}$ is used in the offline phase to determine the reduced model by computing the snapshots $\xi_i := u^{\mathcal{N}}(\mu_i)$ for $\mu_i \in \mathcal{D}$, $1 \leq i \leq N$, and setting $X_N := \text{span}\{\xi_i : 1 \leq i \leq N\}$. The choice of the snapshot samples μ_i is also based upon an error estimator of the form (4), but restricted to the detailed space $\mathcal{X}^{\mathcal{N}} \subsetneq \mathcal{X}$, i.e.,

$$R_N^{\mathcal{N}}(\mu) := \|r(u_N(\mu); \mu)\|_{(\mathcal{X}^{\mathcal{N}})'} = \sup_{v^{\mathcal{N}} \in \mathcal{X}^{\mathcal{N}}} \frac{\langle r(u_N(\mu); \mu), v^{\mathcal{N}} \rangle}{\|u_N(\mu)\|_{\mathcal{X}}}. \tag{12}$$

This is nothing else than computing the Riesz representations in (11) on the truth space $\mathcal{X}^{\mathcal{N}}$, i.e., determine an approximation $\hat{r}_{q,i}^{\mathcal{N}} \in \mathcal{X}^{\mathcal{N}}$ of $\hat{r}_{q,i}$ as

$$\left(\hat{r}_{q,i}^{\mathcal{N}}, v^{\mathcal{N}}\right)_{\mathcal{X}} = \langle r_q(\xi_i), v^{\mathcal{N}} \rangle \qquad \forall v^{\mathcal{N}} \in \mathcal{X}^{\mathcal{N}}. \tag{13}$$

Having these, one computes the corresponding approximation $\boldsymbol{R}_{(q,i),(q',i')}^{\mathcal{N}}$ of $\boldsymbol{R}_{(q,i),(q',i')}$ in (10) and insert this into (9) yielding the approximation $R_N^{\mathcal{N}}(\mu)$ of $R_N(\mu)$. This has an obvious consequence, namely that this does not yield a bound for the exact error as in (5) but only for the *truth error*, i.e.,

$$\|u^{\mathcal{N}}(\mu) - u_N(\mu)\|_{\mathcal{X}} \leq \frac{1}{\alpha(\mu)} R_N^{\mathcal{N}}(\mu) =: \Delta_N^{\mathcal{N}}(\mu).$$

2.4 Estimating the Exact Error

From the above derivation, it should be clear that using a common uniform truth will not yield a control of the exact error. In [8], it was suggested to solve (11) adaptively in order reach any desired accuracy. It turned out, however, that due to the sum in (9), an adaptive error control in (11) is not sufficient for an (online-)efficiently computable error estimate. There are alternative approaches using e.g. an adaptive basis generation, [10, 11]. To the best of our knowledge, however, these approaches are limited to specific problem classes. In this paper, we present an alternative approach using (multi-)wavelets.

3 Wavelet-Based Error Estimation

We start by reviewing the essentials of wavelet bases that are relevant for the problem at hand. For details we refer to [9] and references therein.

3.1 Wavelet Bases

For simplicity, we restrict ourselves to the univariate case and define the scaled and shifted version of a continuous function $g \in C(\mathbb{R})$ of compact support as (j is a *scaling* or *level*, k is a *shift* or the *location* in space) $g_{j,k}(x) := 2^{j/2} g(2^j x - k)$ for $x \in \mathbb{R}$ and $j, k \in \mathbb{Z}$. A system $\Psi := \{\psi_{j,k} : j, k \in \mathbb{Z}\}$ is called *wavelet system* with *d vanishing moments* and *regularity* $s \in \mathbb{R}_+$ if (1) Ψ is a Riesz basis for $L_2(\mathbb{R})$; (2) $\int_{\mathbb{R}} x^p \, \psi_{j,k}(x)\, dx = 0$ for all $j, k \in \mathbb{Z}$ and $0 \le p \le d-1$; (3) $\psi \in H^s(\mathbb{R})$; (4) $|\text{supp}\, \psi_{j,k}| \sim 2^j$. The function ψ is called *mother wavelet*. It is a remarkable fact, [3, 9], that Ψ is not only a Riesz basis for $L_2(\mathbb{R})$ but allows for a characterization of Sobolev spaces, i.e.,

$$\Big\| \sum_{j,k\in\mathbb{Z}} d_{j,k} \, , \psi_{j,k} \Big\|^2_{H^\sigma(\mathbb{R})} \sim \sum_{j,k\in\mathbb{Z}} 2^{2\sigma j} |d_{j,k}|^2, \qquad \sigma \in (0, \min\{s, d\}).^4$$

Moreover, the Riesz representation theorem ensures the existence of a *dual* wavelet basis $\tilde{\Psi} := \{\tilde{\psi}_{j,k} : j, k \in \mathbb{Z}\}$, which is also a wavelet system (for certain parameters $\tilde{d}$, $\tilde{s}$) and $(\psi_{j,k}, \tilde{\psi}_{j',k'})_{L_2(\mathbb{R})} = \delta_{j,j'}\, \delta_{k,k'}$, $j, j', k, k' \in \mathbb{Z}$. The collection $(\Psi, \tilde{\Psi})$ of both wavelet systems is called a *biorthogonal wavelet system*. Examples include biorthogonal B-spline wavelets and orthonormal multi-wavelets, which we use here.

[4] Here $A \sim B$ abbreviates $cA \le B \le CB$ with constants $0 < c \le C < \infty$.

The dual wavelet system also allows a characterization of Sobolev spaces $H^{\sigma}(\mathbb{R})$, $-\min\{\tilde{s}, \tilde{d}\} < \sigma \leq 0$. Note, that these are Sobolev spaces of *negative* order, i.e., *dual* spaces (i.e., those spaces the residual resides in).

There are generalizations for wavelet systems for $L_2(\Omega)$, for bounded domains $\Omega \subset \mathbb{R}^d$. To fix notation, we collect shift and scaling index $\lambda = (j, k)$, denote the level by $|\lambda| := j$ and the index range by $\mathscr{J}$. Then, the relevant dual system takes the form $\tilde{\Psi} := \{\tilde{\psi} : \lambda \in \mathscr{J}\}$ and the norm equivalence reads for $-s < \sigma \leq 0$ with $\boldsymbol{d} := (d_\lambda)_{\lambda \in \mathscr{J}}$ and $\boldsymbol{D} := \operatorname{diag}(2^{|\lambda|})$

$$\|\boldsymbol{d}^T \tilde{\Psi}\|^2_{H^{\sigma}(\Omega)} = \Big\| \sum_{\lambda \in \mathscr{J}} d_\lambda \tilde{\psi}_\lambda \Big\|^2_{H^{\sigma}(\Omega)} \sim \sum_{\lambda \in \mathscr{J}} 2^{2|\lambda|\sigma} |d_\lambda|^2 =: \|\boldsymbol{D}^{\sigma} \boldsymbol{d}\|^2_{\ell_2(\mathscr{J})}. \tag{14}$$

3.2 Wavelet-Based Residual Expansion

The new item presented in this paper is to use (14) for (9). The point of departure is (7). We expand $r_q(w) \in \mathscr{X}'$ by $\tilde{\Psi}$, i.e., $r_q(w) = \sum_{\lambda \in \mathscr{J}} \langle r_q(w), \psi_\lambda \rangle \, \tilde{\psi}_\lambda$, $w \in \mathscr{X}$, and for ξ_i, $1 \leq i \leq N$, we set $d_{(q,i),\lambda} := \langle r_q(\xi_i), \psi_\lambda \rangle$. Then, by (7) and (8)

$$\begin{aligned} r(u_N(\mu); \mu) &= \sum_{q=1}^{Q_r} \vartheta_q^r(\mu) \, r_q(u_N(\mu)) = \sum_{q=1}^{Q_r} \sum_{i=1}^{N} \sum_{\lambda \in \mathscr{J}} \vartheta_q^r(\mu) \, c_i(\mu) \, d_{(q,i),\lambda} \, \tilde{\psi}_\lambda \\ &=: \sum_{\lambda \in \mathscr{J}} \Big(\sum_{m \in M} \sigma_m(\mu) \, d_{m,\lambda} \Big) \tilde{\psi}_\lambda =: \sum_{\lambda \in \mathscr{J}} r_\lambda(\mu) \, \tilde{\psi}_\lambda, \end{aligned}$$

where $M := \{1, \ldots, Q_r\} \times \{1, \ldots, N\}$, $m := (q, i)$ and $\sigma_m(\mu) := \vartheta_q^r(\mu) \, c_i(\mu)$. In order to estimate $R_N(\mu) := \|r(u_N(\mu); \mu)\|_{\mathscr{X}'}$, we have

$$\begin{aligned} R_N(\mu)^2 &\sim \sum_{\lambda \in \mathscr{J}} 2^{-2|\lambda|} \, |r_\lambda(\mu)|^2 = \sum_{\lambda \in \mathscr{J}} 2^{-2|\lambda|} \Big| \sum_{m \in M} \sigma_m(\mu) \, d_{m,\lambda} \Big|^2 \\ &= \sum_{\lambda \in \mathscr{J}} 2^{-2|\lambda|} \sum_{m,m' \in M} \sigma_m(\mu) \, \sigma_{m'}(\mu) \, d_{m,\lambda} \, d_{m',\lambda} \\ &= \sum_{m,m' \in M} \sigma_m(\mu) \, \sigma_{m'}(\mu) \sum_{\lambda \in \mathscr{J}} 2^{-2|\lambda|} d_{m,\lambda} \, d_{m',\lambda} \\ &=: \sum_{m,m' \in M} \sigma_m(\mu) \, \sigma_{m'}(\mu) s_{m,m'} =: R_N^{\Psi}(\mu)^2, \end{aligned} \tag{15}$$

where we abbreviate $s_{m,m'} := \sum_{\lambda \in \mathscr{J}} 2^{-2|\lambda|} d_{m,\lambda} \, d_{m',\lambda}$. These terms are parameter-independent and can be precomputed offline at any desired accuracy, which can be seen as follows. Since $r_q(\xi) \in \mathscr{X}'$, we have $(2^{-|\lambda|} d_{m,\lambda})_{\lambda\in\mathscr{J}} \in \ell_2(\mathscr{J})$, which means that there is a decay with respect to the level. In fact, the following dual version of a Whitney-type estimate is well-known

$$|d_{m,\lambda}| = |\langle r_q(\xi_i), \psi_\lambda\rangle| \leq C\, 2^{-\sigma|\lambda|}\, \|r_q(\xi_i)\|_{H^\sigma(\operatorname{supp}\psi_\lambda)}, \quad -d < \sigma \leq 0, \tag{16}$$

with d as described in Sect. 3.1. This means that we have an a priori estimate for the size of each $d_{m,\lambda}$ in terms of the (known) local Sobolev regularity, which in turn implies that the sum over infinitely many terms in $s_{m,m'}$ can be truncated a priori to a finite one with an a priori bound. This shows that we can indeed precompute $s_{m,m'}$ at any desired tolerance. Finally, the corresponding error estimator reads $\Delta_N^\Psi(\mu) := (c_\Psi\, \alpha(\mu))^{-1} R_N^\Psi(\mu)$, where c_Ψ is the lower equivalence constant in (14).

4 Numerical Experiments

In this section, we compare the numerical performance of the wavelet-based error estimator $\Delta_N^\Psi(\mu)$ in (15) with the standard RB-truth-based $\Delta_N^{\mathscr{N}}(\mu)$ in (12) (implemented in RB-Matlab, [4]) and the exact error. For a fair comparison in terms of CPU time and accuracy, we need access to an exact solution. For this purpose, we consider the simple Laplace problem on $\Omega = (0,1)^2$ with a parameter-dependent source $f(\mu) \in H^{-1}$, i.e., for $x = (x_1, x_2) \in \Omega$, $\mu = (\mu_1, \mu_2) \in \mathscr{D} := [0.2, 0.8]^2$, $a := 1/35$, we set

$$f(x;\mu) := -e^{-\frac{(x_1-\mu_1)^2+(x_2-\mu_2)^2}{a^2}} \left(\frac{4}{a^4}(x_1-\mu_1)^2 + \frac{2}{a^2}(x_2-\mu_2)^2 - \frac{4}{a^4}\right),$$

and let $u(\mu)$ be the corresponding exact solution, which can be computed analytically. Even though $u(\mu) \in C^\infty(\Omega)$, it has a steep gradient (at μ-dependent locations) such that many basis functions are necessary to resolve local details sufficiently well for all $\mu \in \mathscr{D}$.[5] We compare 3 scenarios, namely (1) a truth discretization consisting of 37,249 cubic finite elements (realizing a snapshot tolerance of $\varepsilon = 10^{-2}$); (2) the same with 2,362,369 elements (with $\varepsilon = 10^{-4}$) and (3) an adaptive cubic multi-wavelet snapshot generation as in [1] (where we set the tolerance to 10^{-5}, which leads to 80,637 wavelets at most). In all cases, the RB space X_N of dimension $N = 6$ is computed by a weak greedy on $\mathscr{D}^{\text{train}} := \{0.2, 0.5, 0.8\}^2$ with snapshot orthonormalization.

[5] One might argue that $f(\cdot;\mu)$ is extremely smooth and that the μ-dependence only enters through the right-hand side. Of course, the wavelet-based error estimator equally works in other situations as well. However, we want to do a comparison with the standard RB setting of a common truth. In order to do so, we need (1) a formula for the exact solution, (2) a parameter-dependence which causes local effects. For more complex situations, an even larger improvement is to be expected.

Table 1 Average error for snapshot (S_N) and non-snapshot ($\mathscr{D} \setminus S_N$) parameters over $\mathscr{D}^{\text{test}}$. Case 1: $\mathscr{X}^{\mathscr{N}}$ is cubic FEM grid with 37,249 dof; Case 2: $\mathscr{X}^{\mathscr{N}}$ is cubic FEM grid with 2,362,369 dof; Case 3: Adaptive snapshots based upon cubic multi-wavelets with at most 80,637 dof

∅ error	Case	Exact $\|u(\mu) - u_N(\mu)\|_{\mathscr{X}}$	Truth $\|u^{\mathscr{N}}(\mu) - u_N(\mu)\|_{\mathscr{X}}$	$\Delta_N^{\Psi}(\mu)$	$\Delta_N^{\mathscr{N}}(\mu)$
S_N	1	$9.6 \cdot 10^{-3}$	$2.7 \cdot 10^{-15}$	–	0
	2	$9.2 \cdot 10^{-5}$	$1.6 \cdot 10^{-13}$	–	0
	3	$2.2 \cdot 10^{-6}$	–	$3.6 \cdot 10^{-6}$	–
$\mathscr{D} \setminus S_N$	1	$1.8 \cdot 10^{0}$	$1.8 \cdot 10^{0}$	–	$3.5 \cdot 10^{0}$
	2	$1.8 \cdot 10^{0}$	$1.8 \cdot 10^{0}$	–	$3.5 \cdot 10^{0}$
	3	$1.7 \cdot 10^{0}$	–	$3.1 \cdot 10^{0}$	–

The results are summarized in Table 1. We report average errors over the test set $\mathscr{D}^{\text{test}} := \{0.2, 0.3, \dots, 0.8\}^2 \subset \mathscr{D}$. Of course, in case of a truth discretization, the *truth error* $\|u^{\mathscr{N}}(\mu) - u_N(\mu)\|_{\mathscr{X}}$ and hence the standard RB error estimator $\Delta_N^{\mathscr{N}}(\mu)$ is zero (or machine accuracy) for snapshot parameters $\mu \in S_N$. The *exact error* $\|u(\mu) - u_N(\mu)\|_{\mathscr{X}}$ does not vanish but is in the order of the approximation tolerance of the truth space $\mathscr{X}^{\mathscr{N}}$. For the non-snapshot parameters we observe an effectivity of about 1.94 – as compared to the truth error. The exact error is in the same range and shows that even a highly resolved FE mesh is not capable to guarantee a sufficient accuracy for all parameters.

In case (3), using adaptive snapshot computation, the tolerance is chosen a-priori, which is achieved both by the exact error and the error estimator. This is a bound for the exact error rather than for the truth error (and with much fewer d.o.f.). We obtain an effectivity of about 1.8, i.e., in the same order as in the truth case, but now for the exact error. The online complexity for the wavelet-based estimator $\Delta_N^{\Psi}(\mu)$ is $\mathscr{O}(N^2 + N^2 Q_a^2 + N Q_f + Q_f^2)$ (with Q_a, Q_f as in (6) and $Q_a = 1$, $Q_f = 49$ in this case, $N = 6$), which is online efficient. The CPU time for the wavelet-based error estimator is even too small to measure it with standard tools.[6] We conclude that the wavelet-based error estimator is both theoretically and practically efficient, reliable and effective. Moreover, it allows a control of the exact error.

References

1. M. Ali, K. Steih, K. Urban, *Reduced Basis Methods Based Upon Adaptive Snapshot Computations* (2014). arXiv:1407.1708
2. M. Barrault, Y. Maday, N.C. Nguyen, A.T. Patera, An “empirical interpolation” method: application to efficient reduced-basis discretization of partial differential equations. CR Acad. Sci. Paris I **339**(9), 667–672 (2004)

[6] The computation is finished before C++’s `ctime std::clock function` manages to update the number of clocks.

3. W. Dahmen, Wavelet and multiscale methods for operator equations. Acta Numer. **6**, 55–228 (Cambridge University Press 1997)
4. M. Drohmann, B. Haasdonk, M. Ohlberger, Rbmatlab (2013). http://www.ians.uni-stuttgart.de/MoRePaS/software/rbmatlab
5. J. Hesthaven, G. Rozza, B. Stamm, *Certified Reduced Basis Methods for Parametrized Partial Differential Equations* (Springer, Cham, 2016)
6. D. Huynh, G. Rozza, S. Sen, A. Patera, A successive constraint linear optimization method for lower bounds of parametric coercivity and inf-sup stability constants. CR Math. Acad. Sci. Paris **345**(8), 473–478 (2007)
7. A. Quarteroni, A. Manzoni, F. Negri, *Reduced Basis Methods for Partial Differential Equations*. Unitext, vol. 92 (Springer, Cham, 2016)
8. K. Steih, Reduced basis methods for time-periodic parametric partial differential equations. Ph.D. thesis, Ulm University (2014)
9. K. Urban, *Wavelet Methods for Elliptic Partial Differential Equations* (Oxford University Press, Oxford, 2009)
10. M. Yano, A reduced basis method with exact-solution certificates for symmetric coercive equations. Comput. Meth. Appl. Mech. Eng. **287**, 290–309 (2015)
11. M. Yano, A minimum-residual mixed reduced basis method: exact residual certification and simultaneous finite-element reduced-basis refinement. Math. Model. Numer. Anal. **50**(1), 163–185 (2015)

Model Order Reduction for Pattern Formation in FitzHugh-Nagumo Equations

Bülent Karasözen, Murat Uzunca, and Tuğba Küçükseyhan

Abstract We developed a reduced order model (ROM) using the proper orthogonal decomposition (POD) to compute efficiently the labyrinth and spot like patterns of the FitzHugh-Nagumo (FNH) equation. The FHN equation is discretized in space by the discontinuous Galerkin (dG) method and in time by the backward Euler method. Applying POD-DEIM (discrete empirical interpolation method) to the full order model (FOM) for different values of the parameter in the bistable nonlinearity, we show that using few POD and DEIM modes, the patterns can be computed accurately. Due to the local nature of the dG discretization, the POD-DEIM requires less number of connected nodes than continuous finite element for the nonlinear terms, which leads to a significant reduction of the computational cost for dG POD-DEIM.

1 Introduction

There has been significant development in the efficient implementation and analysis of the model order reduction (MOR) techniques for parametrized partial differential equations (PDEs) [4]. Even though the POD is a very successful MOR technique for linear problems, for nonlinear problems the computational complexity of the evaluation of the nonlinear reduced model still depends on the dimension of the FOM. Several methods are developed to reduce the computational cost so that the nonlinear function evaluations are independent of the dimension of the FOM and the computational complexity is proportional to the dimension of ROM. The discrete empirical interpolation method (DEIM) [3] which is the modified version of the empirical interpolation method (EIM) [2] are the most frequently used ones. The

B. Karasözen (✉) • T. Küçükseyhan
Department of Mathematics and Institute of Applied Mathematics, Middle East Technical University, 06800 Ankara, Turkey
e-mail: bulent@metu.edu.tr; guney.tugba@metu.edu.tr

M. Uzunca
Institute of Applied Mathematics, Middle East Technical University, 06800 Ankara, Turkey
e-mail: uzunca@gmail.com

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_35

DEIM was originally developed for nonlinear functions which depend componentwise on single variables, arising from the finite difference discretization of nonlinear PDEs [3]. When the nonlinear functions are discretized by finite elements, the discretized nonlinear functions depend on the mesh and on the polynomial degree of the finite elements. Therefore the efficiency of the POD-DEIM can be degraded. In [1] the DEIM was applied at different stages of the finite element assembly process. Using the unassembled finite elements, where each DEIM point is related to a single element, the number of nonlinear function calls during the online computation is reduced, but the size of the nonlinear snapshots is enlarged, which increases the offline computational cost [1]. In this paper we consider the dG discretization for time dependent parametrized semi-linear parabolic PDEs. Due to the local nature of the dG approximation, each component of the discretized nonlinear vectors depends only on few elements in the local mesh, whereas the continuous FEMs discretized nonlinear vectors depend on multiple components of the finite element solutions. Therefore the number of POD-DEIM function evaluations for dG approximation is comparable with the finite difference discretization.

In this paper we consider the diffusive parametrized FHN equations [6] with fast diffusing inhibitor v, i.e. we have $D_v > D_u$,

$$\frac{\partial u}{\partial t} = D_u \Delta u - \alpha(v-u) - f(u;\mu), \quad \frac{\partial v}{\partial t} = D_v \Delta u - \beta(v-u), \tag{1}$$

on a space-time cylinder $\Omega \in \mathbb{R}^2 \times (0,T]$ with homogeneous (zero-flux) Neumann boundary conditions. The variables $u(x,t;\mu)$ and $v(x,t;\mu)$ stand for the activator and inhibitor, respectively. The term $f(u;\mu) = (u-\mu)(u^2-1)$ represents the bistable nonlinearity for the parameter μ. We investigate the formation of labyrinth and spot like patterns for different values of the parameter μ as in [6], where α, β and diffusion coefficients D_u, D_v are fixed.

The paper is organized as follows. In Sect. 2 the discretization of FHN equation (1) in space by symmetric interior penalty discontinuous Galerkin method (SIPG) is given. In Sect. 3 the ROM based on the POD is formulated. In Sect. 4 we describe the SIPG discretized version of the DEIM for the bistable nonlinearity. Numerical results for pattern formations for different values of the parameters demonstrate the good performance of dG ROMs. The paper ends with some conclusions and outlook for the future work.

2 Full Order Model

The FHN equation (1) is discretized in space using SIPG method [7]. Let ε_h be the disjoint partition of the domain $\Omega \subset \mathbb{R}^2$ with elements (triangles) $\{E_i\}_{i=1}^{N_{el}} \in \varepsilon_h$, where N_{el} is the number of elements in the partition. The discrete solution and test function spaces on ε_h are given by

$$D_q = D_q(\varepsilon_h) := \{\vartheta \in L^2(\Omega) : \vartheta_E \in \mathbb{P}_q(E),\ \forall E \in \varepsilon_h\},$$

where $\mathbb{P}_q(E)$ is the space of polynomials of degree at most q on $E \in \varepsilon_h$, and the functions $\vartheta \in D_q$ are discontinuous along the inter-element boundaries. Multiplying (1) by arbitrary test functions $\vartheta_1, \vartheta_2 \in D_q$ and integrating by using Green's theorem over each mesh element, we obtain the semi-discrete variational equations

$$\begin{aligned}\left(\frac{\partial u_h}{\partial t}, \vartheta_1\right) + a_h(D_u; u_h, \vartheta_1) + \alpha(v_h - u_h, \vartheta_1) + (f(u_h; \mu), \vartheta_1) &= 0,\\ \left(\frac{\partial v_h}{\partial t}, \vartheta_2\right) + a_h(D_v; u_h, \vartheta_2) + \beta(v_h - u_h, \vartheta_2) &= 0,\end{aligned} \tag{2}$$

where $a_h(D_u; u, \vartheta_1)$ and $a_h(D_v; u, \vartheta_2)$ stand for the dG bilinear forms given by [7]

$$\begin{aligned}a_h(D; w, \vartheta) = &\sum_{E\in\varepsilon_h} \int_E D\nabla w \cdot \nabla\vartheta - \sum_{e\in\Gamma_h^0} \int_e \{D\nabla w\}[\vartheta]\mathrm{d}s\\ &- \sum_{e\in\Gamma_h^0} \int_e \{D\nabla\vartheta\}[w]\mathrm{d}s + \sum_{e\in\Gamma_h^0} \frac{\kappa D}{h_e} \int_e [w][\vartheta]\mathrm{d}s,\end{aligned}$$

where κ is called the penalty parameter depending only on the polynomial order q, see [7] for details.

Let $N := N_{loc} \times N_{el}$ denotes the dG degrees of freedom (DoFs), where N_{loc} is the local dimension on each element depending on the polynomial degree q. Then, for any $t \in (0, T]$, the dG solutions of (2) are of the form

$$u_h(t, x) = \sum_{i=1}^{N} u_i(t)\phi_i(x) = \phi\mathbf{u}\ , \quad v_h(t, x) = \sum_{i=1}^{N} v_i(t)\phi_i(x) = \phi\mathbf{v}, \tag{3}$$

where $\mathbf{u}(t) := (u_1(t), \ldots, u_N(t))^T$ and $\mathbf{v}(t) := (v_1(t), \ldots, v_N(t))^T$ are the vectors of time dependent unknown coefficients of u_h and v_h, respectively, and $\phi(x) := [\phi_1(x)\ \ \ldots\ \ \phi_N(x)]$ is the matrix of the basis functions. Plugging (3) into the equations (2) and choosing $\vartheta_1 = \vartheta_2 = \phi_i$, $i = 1, \cdots, N$, we obtain the FOM of (1) as the following system of ordinary differential equations (ODEs)

$$\begin{aligned}M\mathbf{u}_t + S_u\mathbf{u} + \alpha M(\mathbf{v} - \mathbf{u}) + F(\mathbf{u}; \mu) &= 0,\\ M\mathbf{v}_t + S_v\mathbf{v} + \beta M(\mathbf{v} - \mathbf{u}) &= 0,\end{aligned} \tag{4}$$

where $S_u, S_v \in \mathbb{R}^{N\times N}$ are the stiffness matrices, $M \in \mathbb{R}^{N\times N}$ is the mass matrix and $F(\mathbf{u};\mu) \in \mathbb{R}^N$ is the nonlinear vector depending on the parameter μ. The FOM (4) is solved by the backward Euler method.

3 Reduced Order Model

For an arbitrary parameter $\bar{\mu}$, the k-th order approximate ROM solutions $\tilde{u}_{h,k} := \tilde{u}_{h,k}(t,x;\bar{\mu})$ and $\tilde{v}_{h,k} := \tilde{v}_{h,k}(t,x;\bar{\mu})$ have the form

$$\tilde{u}_{h,k} = \sum_{i=1}^{k} \tilde{u}_i(t)\psi_{u,i}(x) \ , \quad \tilde{v}_{h,k} = \sum_{i=1}^{k} \tilde{v}_i(t)\psi_{v,i}(x), \tag{5}$$

where $\tilde{\mathbf{u}}(t) := (\tilde{u}_1(t),\dots,\tilde{u}_k(t))^T$ and $\tilde{\mathbf{v}}(t) := (\tilde{v}_1(t),\dots,\tilde{v}_k(t))^T$ are the coefficient vectors of the ROM solutions. For a set $\{\mu_1,\dots,\mu_{n_s}\}$ of parameter samples, the POD reduced basis functions $\{\psi_{u,i}\}$ and $\{\psi_{v,i}\}$ are computed as solutions of the minimization problem

$$\begin{aligned}&\min_{\psi_{w,1},\dots,\psi_{w,k}} \frac{1}{n_s}\sum_{m=1}^{n_s}\frac{1}{J}\sum_{j=1}^{J}\left\| w^{m,j} - \sum_{i=1}^{k}(w^{m,j},\psi_{w,i})_{L^2(\varepsilon_h)}\psi_{w,i}\right\|^2_{L^2(\varepsilon_h)} \\ &\text{subject to } (\psi_{w,i},\psi_{w,j})_{L^2(\varepsilon_h)} = \Psi^T_{w,\cdot,i}M\Psi_{w,\cdot,j} = \delta_{ij} \ , \ 1 \le i,j \le k,\end{aligned} \tag{6}$$

where $w^{m,j} \approx w(t_j,x;\mu_m)$ denotes the approximate solution at the time t_j for a fixed parameter μ_m, for $w \in \{u,v\}$, $m = 1,\dots,n_s$, and δ_{ij} is the Kronecker delta. We note that $w^{m,j}$ in (6) stands for the solution but not for the coefficient vector of the unknown solution as in the continuous finite elements. For the dG discretizations we use modal basis functions where the coefficients do not coincide with the solution values. Therefore, instead of the Euclidean norm, we use in (6) the weighted inner product and the corresponding norm with the symmetric positive definite mass matrix M, leading to M-orthogonal reduced basis functions [5].

In practice, instead of the minimization problem (6), an equivalent eigenvalue problem is solved [5].

$$\widehat{\mathscr{U}}\widehat{\mathscr{U}}^T\widehat{\Psi}_{u,\cdot,i} = \sigma^2_{u,i}\widehat{\Psi}_{u,\cdot,i} \ , \qquad \widehat{\mathscr{V}}\widehat{\mathscr{V}}^T\widehat{\Psi}_{v,\cdot,i} = \sigma^2_{v,i}\widehat{\Psi}_{v,\cdot,i} \ , \quad i = 1,2,\dots,k, \tag{7}$$

where $\widehat{\mathscr{U}} = R\mathscr{U}$, $\widehat{\mathscr{V}} = R\mathscr{V}$, $\widehat{\Psi}_{\cdot,\cdot,i} = R\Psi_{\cdot,\cdot,i}$, R^T is the Cholesky factor of the mass matrix M, and $\mathscr{U} = [\mathbf{u}^{1,1},\dots,\mathbf{u}^{n_s,J}]$ and $\mathscr{V} = [\mathbf{v}^{1,1},\dots,\mathbf{v}^{n_s,J}]$ in $\mathbb{R}^{N\times(n_s\times J)}$ are the snapshot matrices. The vectors $\Psi_{u,\cdot,i}$ and $\Psi_{v,\cdot,i}$ denote the coefficient vectors of the reduced basis functions $\psi_{u,i}$ and $\psi_{v,i}$, respectively. The solutions $\widehat{\Psi}_{\cdot,\cdot,i}$ of (7) are obtained as the first k left singular vectors in the generalized singular value decomposition (SVD) of $\widehat{\mathscr{U}}$ and $\widehat{\mathscr{V}}$, respectively [5]. Combining the FOM (3) and

ROM (5) solutions and using the fact that the reduced basis functions $\psi_{u,i}$ and $\psi_{v,i}$ belong to the dG space D_q, we obtain the relations: $\mathbf{u} = \Psi_u \tilde{\mathbf{u}}$ and $\mathbf{v} = \Psi_v \tilde{\mathbf{v}}$, between the coefficient vectors $\mathbf{u}$, $\mathbf{v}$ of the FOM solutions and the coefficient vectors $\tilde{\mathbf{u}}$, $\tilde{\mathbf{v}}$ of the ROM solutions. Substituting these relations into (4) and projecting onto the reduced spaces spanned by $\{\psi_{u,1}, \ldots, \psi_{u,k}\}$ and $\{\psi_{v,1}, \ldots, \psi_{v,k}\}$, respectively, we obtain the k-dimensional ROM

$$\begin{aligned} \tilde{\mathbf{u}}_t + \tilde{S}_u \tilde{\mathbf{u}} + \alpha \tilde{M}_u \tilde{\mathbf{v}} - \alpha \tilde{\mathbf{u}} + \Psi_u^T F(\Psi_u \tilde{\mathbf{u}}; \bar{\mu}) = 0, \\ \tilde{\mathbf{v}}_t + \tilde{S}_v \tilde{\mathbf{v}} + \beta \tilde{\mathbf{v}} - \beta \tilde{M}_v \tilde{\mathbf{u}} = 0 \end{aligned} \tag{8}$$

with the reduced matrices

$$\tilde{S}_u = \Psi_u^T S_u \Psi_u \ , \quad \tilde{S}_v = \Psi_v^T S_v \Psi_v \ , \quad \tilde{M}_u = \Psi_u^T M \Psi_v \ , \quad \tilde{M}_v = \Psi_v^T M \Psi_u .$$

The system (8) is solved by the backward Euler method, as well.

4 Discrete Empirical Interpolation Method (DEIM)

Although the dimension of the reduced system (8) is small, $k \ll N$, the computation of the nonlinear term $N(\tilde{\mathbf{u}}) := \Psi_u^T F(\Psi_u \tilde{\mathbf{u}}; \bar{\mu})$ still depends on the dimension N of the full system. We apply DEIM [3] to reduce the computational cost, where the nonlinear function is approximated as $F(\Psi_u \tilde{\mathbf{u}}; \bar{\mu}) \approx Ws(t)$, from a subspace $W = [W_1, \ldots, W_n] \in \mathbb{R}^{N \times n}$, where each member W_i is called the DEIM basis functions, $i = 1, 2, \ldots, n$ ($n \ll N$). The DEIM basis functions W_i are computed through the SVD of the nonlinear snapshot matrix $\mathscr{F} := [F^{1,1}, \ldots, F^{n_s,J}] \in \mathbb{R}^{N \times (n_s \times J)}$, where $F^{m,i} := F(\Psi_u \tilde{\mathbf{u}}(t_i); \mu_m)$ are the nonlinear vectors at the time instance t_i, obtained in the online computation for the parameters μ_m, $m = 1, \ldots, n_s$. Because the system $Ws(t)$ is overdetermined, the projection matrix P is introduced which is computed by the greedy DEIM algorithm [3]. Then, we use the approximation $N(\tilde{\mathbf{u}}) \approx \tilde{N}(\tilde{\mathbf{u}}) = Q\tilde{F}^{\bar{\mu}}$ where the matrix $Q = \Psi_u^T W (P^T W)^{-1} \in \mathbb{R}^{k \times n}$ is precomputable and $\tilde{F}^{\bar{\mu}} = P^T F(\Psi_u \tilde{\mathbf{u}}; \bar{\mu}) \in \mathbb{R}^n$ is the n-dimensional non-linear vector which can be computed in an efficient way. In addition, the DEIM approximation satisfies the a priori error bound

$$\| F^{\bar{\mu}} - W(P^T W)^{-1} \tilde{F}^{\bar{\mu}} \|_2 \leq \| (P^T W)^{-1} \|_2 \| (I - WW^T) F^{\bar{\mu}} \|_2,$$

where the term $\| (P^T W)^{-1} \|_2$ is of moderate size of order 100 or less [1].

In dG discretization, the integrals are computed on a single triangular element, whereas for continuous finite element discretizations with linear polynomials all the interior degrees of freedoms are shared by 6 triangular elements, see Fig. 1. The unassembled finite element approach is used in [1], so that each DEIM point is related to one element, which reduces the online computational cost, but increases the number of snapshots and therefore the cost of the offline computation. Due to

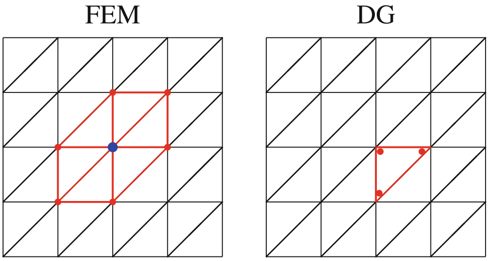


Fig. 1 Connectivity of degrees of freedoms for linear basis functions

its local nature, the dG discretization is automatically in the unassembled form and it does not require computation of additional snapshots.

As we use for the time integration the implicit backward Euler method, on each time step the nonlinear equations have to be solved by Newton's method. Therefore the computational cost of the Jacobian by DEIM of the reduced model has to be taken into account. Because the support of dG basis functions has only one single element, the Jacobian matrices of the FOMs appear in block diagonal form unlike the continuous FEMs where the Jacobian matrices contain overlapping blocks. The Jacobian matrices arising from POD and POD-DEIM are of the form

$$\frac{\partial}{\partial \tilde{\mathbf{u}}} N(\tilde{\mathbf{u}}) = \Psi_u^T J_F^{\bar{\mu}} \Psi_u \;, \quad \frac{\partial}{\partial \tilde{\mathbf{u}}} \tilde{N}(\tilde{\mathbf{u}}) = Q(P^T J_F^{\bar{\mu}}) \Psi_u,$$

where $(P^T J_F^{\bar{\mu}}) \in \mathbb{R}^{n \times N}$ is the matrix including only $n \ll N$ rows of the Jacobian $J_F^{\bar{\mu}}$, and in each row of the Jacobian there are only N_{loc} nonzero terms because of the local structure of the dG. Hence, only $n \times N_{loc}$ entries are needed to compute $P^T J_F^{\bar{\mu}}$, whereas without DEIM, $N_{el} \times N_{loc}^2$ entries are required for computation of the Jacobian $J_F^{\bar{\mu}}$ of the FOM.

5 Numerical Results

We consider FHN equation (1) for $(x,t) \in [-10,10]^2 \times [0,1000]$ with random initial conditions uniformly distributed between -1 and 1. The other parameters $D_u = 0.04$, $D_v = 1$, $\alpha = 0.3$, $\beta = 1$ are fixed as in [6]. We use linear dG polynomials ($N_{loc} = 3$), and as the discrete mesh, we form the partition of $[-10,10]^2$, by 5 times uniform refinement, with 2048 triangular elements leading to 6144 DoFs. Snapshots are taken in the time interval $[0,1000]$ with the time step $\Delta t = 0.5$. For POD/POD-DEIM basis construction, we use the parameter samples $\mu \in \{-0.04, -0.02, 0, 0.02, 0.04\}$, $n_s = 5$. The reduced systems are solved for the set $\{-0.03, -0.01, 0.01, 0.03\}$ of parameter values of μ, which are not contained in the set of sample parameters. The average number of Newton iterations was 1 for the computation of the FOMs and ROMs on each time step.

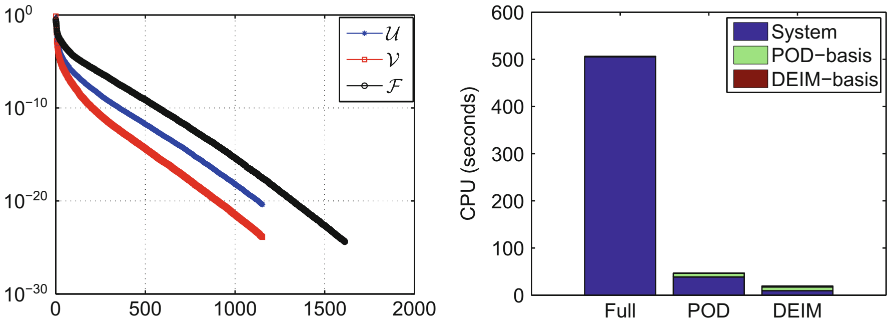


Fig. 2 (*Left*) Decay of the singular values of solution snapshots $\mathscr{U}$, $\mathscr{V}$ and of the nonlinear snapshots $\mathscr{F}$; (*Right*) CPU times for the computation of FOMs, POD and POD-DEIM ROMs for the parameter value $\mu = 0.03$

Table 1 The computation times (in sec), speed-up factors S_{POD} and S_{DEIM}, and the DEIM projection error bounds $\|(P^T W)^{-1}\|_2$

μ	FOM	POD	POD-DEIM	S_{POD}	S_{DEIM}	$\|(P^T W)^{-1}\|_2$
−0.03	527.3	34.5	7.5	15.31	70.21	28
−0.01	501.9	33.4	13.2	15.05	38.08	33
0.01	522.3	32.9	11.9	15.88	43.67	41
0.03	505.9	38.6	9.0	13.10	56.43	33

The decay of the singular values for the solution snapshots $\mathscr{U}$, $\mathscr{V}$ and nonlinear snapshots is given in Fig. 2, left, and the CPU times of the FOMs and ROMs for the parameter value $\mu = 0.03$ are shown in Fig. 2, right. In Table 1 we give the CPU times for FOMs, POD and POD-DEIM ROMs together with the speed-up factors S_{POD} and S_{DEIM}, which demonstrate the efficiency of the DEIM. We note that in the POD-DEIM algorithm, the nonlinearity is discretized at six points of the mesh by linear continuous FEM, whereas at three points of the mesh by linear dG method (see Fig. 1). Therefore, during the online computation, dG requires less more work than the continuous FEM. In Fig. 3, the patterns of FOMs, POD and POD-DEIM reduced solutions are shown at the final time $T = 1000$. The ROM patterns in Fig. 3 computed with POD are very close to those of the FOMs as in [6]. But the patterns computed with POD-DEIM are less accurate than those with the POD computed ones for some parameter values in Fig. 3. The DEIM does not improve the accuracy of the POD reduced model, but enormously reduces the computational complexity [1]. The error bounds $\|(P^T W)^{-1}\|_2$ of moderate size for the DEIM approximations are also given in Table 1.

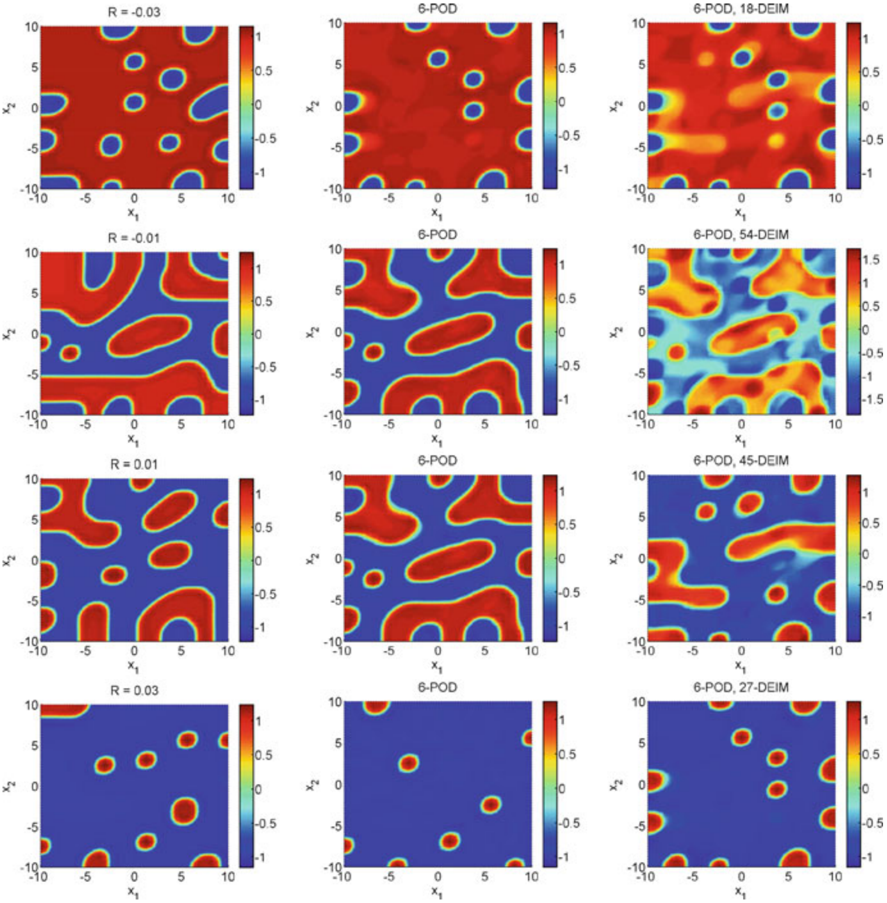


Fig. 3 Patterns for u at the final time $T = 1000$ with FOM (*left*), POD (*middle*) and POD-DEIM (*right*) for the parameter values $\mu \in \{-0.03, -0.01, 0.01, 0.03\}$ from top to bottom

6 Conclusions and Outlook

We have demonstrated that the dG approximation can produce due to its local structure cost effective and accurate reduced order solutions by approximating the parameter dependent nonlinear terms with the DEIM. In a future work we will consider the parametrized FHN equation with the diffusivity coefficients D_u and D_v to compute the reduced order solutions by preserving the multiscale dynamics of the activator u and the inhibitor v in time. Because the size of the SVD problem can be prohibitive for the global POD, we will also apply the greedy POD.

Acknowledgements This work has been supported by METU BAP-07-05-2015 009.

References

1. H. Antil, M. Heinkenschloss, D.C. Sorensen, Application of the discrete empirical interpolation method to reduced order modeling of nonlinear and parametric systems, in *Reduced Order Methods for Modeling and Computational Reduction*, ed. by A. Quarteroni, G. Rozza. Modeling, Simulation and Applications, vol. 9 (Springer, Cham, 2014), pp. 101–136
2. M. Barrault, Y. Maday, N.C. Nguyen, A.T. Patera, An empirical interpolation method: application to efficient reduced-basis discretization of partial differential equations. Comptes Rendus Math. **339**(9), 667–672 (2004)
3. S. Chaturantabut, D.C. Sorensen, Nonlinear model reduction via discrete empirical interpolation. SIAM J. Sci. Comput. **32**, 2737–2764 (2010)
4. J.S. Hesthaven, G. Rozza, B. Stamm, *Certified Reduced Basis Methods for Parametrized Partial Differential Equations*. SpringerBriefs in Mathematics (Springer/Basque Center for Applied Mathematics, Bilbao, 2016)
5. M. Hinze, S. Volkwein, Proper orthogonal decomposition surrogate models for nonlinear dynamical systems: error estimates and suboptimal control, in *Dimension Reduction of Largescale Systems*, ed. by P. Benner, D.C. Sorensen, V. Mehrmann. Lecture Notes in Computational Science and Engineering, vol. 45 (Springer, Berlin/New York, 2005), pp. 261–306
6. T.T. Marquez-Lago, P. Padilla, A selection criterion for patterns in reaction diffusion systems. Theor. Biol. Med. Model. **11**, 1–17 (2014)
7. B. Rivière, *Discontinuous Galerkin Methods for Solving Elliptic and Parabolic Equations: Theory and Implementation* (SIAM, Philadelphia, 2008)

Local Parametrization of Subspaces on Matrix Manifolds via Derivative Information

Ralf Zimmermann

Abstract A method is proposed for constructing local parametrizations of orthogonal bases and of subspaces by computing trajectories in the Stiefel and the Grassmann manifold, respectively. The trajectories are obtained by exploiting sensitivity information on the singular value decomposition with respect to parametric changes and a Taylor-like local linearization suitably adapted to the underlying manifold structure. An important practical application of the proposed approach is parametric model reduction (pMOR). The connection with pMOR is discussed in detail and the results are illustrated by numerical experiment.

1 Motivation: Parametric Model Reduction

The basic objective in model reduction is to emulate a large-scale dynamical system with very few degrees of freedom. While classical model reduction techniques aim at producing an accurate low-order approximation to the time trajectory of the original dynamical system, in parametric model reduction (pMOR) it is tried to account for additional system parameters, e.g. varying operating conditions. We explain the process of projection-based MOR with the aid of a generic example. Consider a spatio-temporal dynamical system in semi-discrete form

$$\frac{\partial}{\partial t}y(t,\mu) = f(y(t,\mu)), \quad y(t_0,\mu) = y_{0,\mu}, \tag{1}$$

where $y(t,\mu) \in \mathbb{R}^n$ is the spatially discretized *state vector* of dimension n, μ denotes additional system parameters or operating conditions and $f : \mathbb{R}^n \to \mathbb{R}^n$ may be nonlinear. Projection-based MOR starts with constructing a suitable low-dimensional subspace that acts as a space of candidate solutions.

R. Zimmermann (✉)
TU Braunschweig, Institute Computational Mathematics, Pockelsstr. 14, D-38106 Braunschweig, Germany
e-mail: ralf.zimmermann@tu-bs.de

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_36

Subspace construction One way to construct the required projection subspace is proper orthogonal decomposition (POD). In its simplest form, POD can be summarized as follows. For a fixed $\mu = \mu_0$, let $y^1 := y(t_1, \mu_0), \dots, y^m := y(t_m, \mu_0) \in \mathbb{R}^n$ be a set of state vectors and let $Y := \left(y^1, \dots, y^m\right) \in \mathbb{R}^{n\times m}$. The state vectors y^i are called *snapshots* and the matrix Y is called the associated *snapshot matrix*. POD is concerned with finding a subspace $\mathscr{U}$ of dimension $p \leq m$ represented by a column-orthogonal matrix $U \in \mathbb{R}^{n\times p}$ such that the error between the input snapshots and their orthogonal projection onto $\mathscr{U}$ is minimized:

$$\min_{U\in\mathbb{R}^{n\times p}, U^TU=I} \sum_k \|y^k - UU^Ty^k\|_2^2 \quad \left(\Leftrightarrow \min_{U\in\mathbb{R}^{n\times p}, U^TU=I} \|Y - UU^TY\|_F^2\right).$$

The main result of POD is that for any $p \leq m$, the best p-dimensional approximation of colspan$\{y^1, \dots, y^m\}$ in the above sense is $\mathscr{U} = $ colspan$\{u^1, \dots, u^p\}$, where $\{u^1, \dots, u^p\}$ are the eigenvectors of the matrix UU^T corresponding to the p largest eigenvalues. The subspace $\mathscr{U}$ is called *POD subspace*. The same subspace is obtained via a singular value decomposition (SVD) of the snapshot matrix $Y = U\Sigma V^T$, truncated to the first $p \leq m$ columns of U. For more details, see, e.g. [3, §3.3]. Since the input snapshots are supplied at a fixed operating condition μ_0, the POD subspace is considered to be a space of solution candidates $\mathscr{U}(\mu_0) = $ colspan$(U(\mu_0))$ valid at μ_0.

Projection POD leads to a parameter decoupling via

$$\tilde{y}(t, \mu_0) = U(\mu_0)y_r(t). \tag{2}$$

In this way, the time trajectory of the reduced model is uniquely defined by the coefficient vector $y_r(t) \in \mathbb{R}^p$ that represents the reduced state vector with respect to the subspace colspan$(U(\mu_0))$. Let $W(\mu_0)$ be such that the matrix pair $U(\mu_0), W(\mu_0)$ is biorthogonal, i.e. $W(\mu_0)^TU(\mu_0) = I$. Inserting (2) in (1) and multiplying with $W(\mu_0)^T$ from the left leads to

$$\frac{d}{dt}y_r(t) = W^T(\mu_0)f(U(\mu_0)y_r(t)). \tag{3}$$

This approach goes by the name of Petrov-Galerkin projection, if $W(\mu_0) \neq U(\mu_0)$ and Galerkin-Projection if $W(\mu_0) = U(\mu_0)$.

There are various ways to proceed from Eq. (3) depending on the nature of the function f, which may be categorized as linear, affine linear or nonlinear, see [3] for a recent survey.

Problem statement The main focus of the work at hand is *not* on the efficient solution of the reduced system (3). Rather, I will address the issue of parameterizing the POD subspaces and their underlying orthonormal bases (ONBs). To this end, we consider the set of column-orthogonal matrices $\{U \in \mathbb{R}^{n\times p} | U^TU = I\} =: St(n, p)$, called the *Stiefel manifold*, and the set of p-dimensional subspaces $\{\mathscr{U} \leq \mathbb{R}^n | \dim(\mathscr{U}) = p\} =: Gr(n,p)$, called the *Grassmann manifold*. Then, the objective is to construct trajectories

$$\mu \mapsto U(\mu) \in St(n,p) \text{ and } \mu \mapsto \mathscr{U}(\mu) = \text{colspan}(U(\mu)) \in Gr(n,p), \qquad (4)$$

respectively. The manifold structure of the Stiefel and the Grassmann manifold will be exploited in order to obtain a Taylor-like first order approximation close to an expansion point μ_0. In view of MOR applications, it is important that the trajectories (4) are obtained *without* computing additional snapshots.

2 A Taylor-Like Expansion for Stiefel and Grassmann Curves

First, let us fix some notation. The *orthogonal group*, i.e., the set of all (p-by-p) square orthogonal matrices is denoted by $O_{p\times p}$. The tangent space in a point $\mathscr{U} \in Gr(n,p)$ is $T_{\mathscr{U}}Gr(n,p) \cong \text{range}(I - \Pi_{\mathscr{U}}) \subset \mathbb{R}^{n\times p}$, where $\Pi_{\mathscr{U}}$ is the orthogonal projection onto $\mathscr{U}$. The tangent space at $U \in St(n,p)$ is $T_U St(n,p) \cong \{\Delta \in \mathbb{R}^{n\times p} | \Delta^T U = -U^T \Delta\}$. In [5], it was shown that any tangent vector $\Delta \in T_U St(n,p)$ has the form $\Delta = UA + (I - UU^T)C$, where $A \in \mathbb{R}^{p\times p}$ is skew-symmetric and $C \in \mathbb{R}^{n\times p}$ is arbitrary. The tangent spaces of the Stiefel manifold and the Grassmannian carry a Riemannian inner product, which is defined by $g_U^{St}(\Delta, \tilde{\Delta}) = tr(\Delta^T(I - \frac{1}{2}UU^T)\tilde{\Delta})$ for $\Delta, \tilde{\Delta} \in T_U St(n,p)$ and $g_{\mathscr{U}}^{Gr}(\Delta, \tilde{\Delta}) = tr(\Delta^T \tilde{\Delta})$ for $\Delta, \tilde{\Delta} \in T_{\mathscr{U}}Gr(n,p)$, see [5, §2.4.2, §2.5.1]. The Grassmann manifold can be realized as a quotient manifold of the Stiefel manifold:

$$Gr(n,p) = St(n,p)/O_{p\times p} = \{[U] | \quad U \in St(n,p)\}. \qquad (5)$$

Hence, by definition, two matrices $U, \tilde{U} \in St(n,p)$ define the same point $[U] \in Gr(n,p)$ if there exists an $R \in O_{p\times p}$ such that $U = \tilde{U}R$. As in [5], we will make use throughout of the quotient representation (5) of the Grassmann manifold with matrices in $St(n,p)$ acting as representatives in numerical computations. For details, the reader is referred to [1, 5].

Geodesics and the Riemannian exponential mapping Consider an arbitrary Riemannian manifold $\mathscr{M}$. Geodesics on $\mathscr{M}$ are locally shortest paths that are parametrized by the arc length. Because they satisfy an initial value problem, they are uniquely defined by specifying a starting point $p_0 \in \mathscr{M}$ and a starting velocity $\Delta \in T_{p_0}\mathscr{M}$. Geodesics give rise to the Riemannian exponential mapping that takes a tangent vector $\Delta \in T_{p_0}\mathscr{M}$ to the endpoint $\mathscr{C}(1)$ of a geodesic path $\mathscr{C} : [0,1] \to \mathscr{M}$ starting at $\mathscr{C}(0) = p_0 \in \mathscr{M}$ with velocity $\Delta = \dot{\mathscr{C}}(0) \in T_{p_0}\mathscr{M}$. It is denoted by $Exp_{p_0} : T_{p_0}\mathscr{M} \to \mathscr{M}, Exp_{p_0}(\Delta) := \mathscr{C}(1)$. The Riemannian exponential is locally a radial-isometric diffeomorphism. Numerically efficient formulae for the

Algorithm 1 [5, §2.4.2] Stiefel exponential as a curve $t \mapsto Exp^{St}_{U_0}(t\Delta)$

```
Require: base point $U_0 \in St(n,p)$, $\Delta \in T_{U_0}St(n,p)$
1: $QR := H = (I - U_0 U_0^T)\Delta \in \mathbb{R}^{n\times p}$          {qr-decomp. of normal component of $\Delta$}
2: $A := U_0^T \Delta \in \mathbb{R}^{p\times p}$                               {horizontal component, skew}
3: $\begin{pmatrix} A & -R^T \\ R & 0 \end{pmatrix} = T\Lambda T^H \in \mathbb{R}^{2p\times 2p}$                                   {EVD}
Ensure: $Exp^{St}_{U_0}(t\Delta) = (U_0, Q)\left(T\exp(t\Lambda)T^H(I_p, 0)^T\right) \in St(n,p)$
    Costs: $4np^2 + \mathcal{O}(p^3)$ for each $t$ when computing 1.–3. a priori (offline).
```

Algorithm 2 [5, §2.5.1] Grassmann exponential as a curve $t \mapsto Exp^{Gr}_{\mathcal{U}_0}(t\Delta)$

```
Require: base point $\mathcal{U}_0 = [U_0] \in Gr(n,p)$, where $U_0 \in St(n,p)$, $\Delta \in T_{\mathcal{U}_0}Gr(n,p)$
1: $\Delta \overset{SVD}{=} Q\Sigma V^T$, with $Q \in St(n,p)$, $\Sigma \in \mathbb{R}^{p\times p}$ diagonal, $V \in O_{p\times p}$.
2: $U(t) = (U_0 V)\cos(t\Sigma) + Q\sin(t\Sigma)$                {cos, sin applied only to diagonal}
Ensure: $Exp^{Gr}_{\mathcal{U}_0}(t\Delta) = [U(t)] \in Gr(n.p)$
    Costs: $2np + \mathcal{O}(p)$ for each $t$ when computing 1. and $U_0 V$ a priori (offline).
```

Riemannian exponential mappings on $St(n,p)$ and $Gr(n,p)$ have been derived in [5, §2.5.1] and are given in Algorithms 1 and 2.

Coming back to the problem statement, we will make use of the Stiefel and Grassmann exponential in order to parametrize trajectories of ONBs and subspaces, respectively. The basic idea is as follows. Suppose that at an operating point μ_0 we have a projection subspace $\mathcal{U}(\mu_0)$ associated with a reduced model that is represented by a matrix $U(\mu_0) \in St(n,p)$. If we can specify an appropriate starting velocity $\Delta \in T_{U(\mu_0)}St(n,p)$, then

$$[0, \delta\mu] \to St(n,p), \mu \mapsto Exp^{St}_{U_0}(\mu\Delta) \qquad (6)$$

is a curve on $St(n,p)$, i.e. a trajectory of ONBs. It turns out that with the choice of $\Delta = \frac{dU}{d\mu}(\mu_0)$, the curve $Exp^{St}_{U_0}(\mu\Delta)$ matches the exact ONB $U(\mu_0 + \mu)$ up to terms of order $\mathcal{O}(\mu^2)$, see Lemmata 1 and 2 below. This procedure is visualized in Fig. 1. Hence, we require a formula for the derivative of $U(\mu)$, which stems from a snapshot POD/SVD.

Remark The geodesic (6) can be considered as a building block of an analytic path of the SVD factorization. The investigations about analytic SVDs have a long tradition, see, e.g. [4] and references therein. In fact, the method introduced in [4] (for the full SVD) eventually leads to an approximation of the geodesic path by means of a numerical minimization of the arc length rather than the closed form solution of the associated initial value problem, which was not available at the time of the writing of [4].

Differentiating the SVD Suppose that $\mu \mapsto Y(\mu) \in \mathbb{R}^{n\times p}$ is a differentiable matrix curve. If the singular values of $Y(\mu_0)$ are mutually distinct, then the singular values and both the left and the right singular vectors depend differentiable on

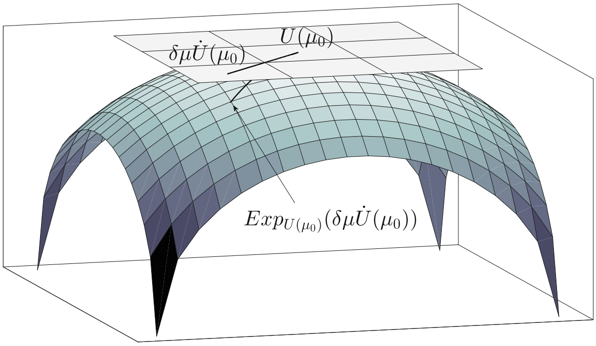

$\delta\mu\dot{U}(\mu_0)$
$U(\mu_0)$
$Exp_{U(\mu_0)}(\delta\mu\dot{U}(\mu_0))$

Fig. 1 Visualization of the geodesic/exponential associated with a tangent vector

$\mu \in [\mu_0 - \delta\mu, \mu_0 + \delta\mu]$ for $\delta\mu$ small enough. For brevity, let $\dot{Y} = \frac{dY}{d\mu}(\mu_0)$ denote the derivative with respect to μ evaluated in μ_0 and so forth. Let $\mu \mapsto Y(\mu) = U(\mu)\Sigma(\mu)V(\mu)^T \in \mathbb{R}^{n\times p}$ and let $C(\mu) = (Y^TY)(\mu)$. Let u^j and v^j, $j = 1, \dots, p$ denote the columns of $U(\mu_0)$ and $V(\mu_0)$, respectively. It holds

$$\dot{\sigma}_j = (u^j)^T\dot{Y}v^j, \tag{7}$$

$$\dot{V} = V\Gamma, \text{ where } \Gamma_{ij} = \begin{cases} \frac{(v^i)^T\dot{C}v^j}{(\sigma_j+\sigma_i)(\sigma_j-\sigma_i)}, & i \neq j \\ 0, & i = j \end{cases} \text{ for } i,j = 1,\dots,p, \tag{8}$$

$$\dot{U} = \dot{Y}V\Sigma^{-1} + Y\dot{V}\Sigma^{-1} + YV\dot{\Sigma}^{-1} = \left(\dot{Y}V + U(\Sigma\Gamma - \dot{\Sigma})\right)\Sigma^{-1}. \tag{9}$$

A proof can be found in [6]. Note that $U^T(\mu_0)\dot{U}(\mu_0)$ is skew-symmetric so that indeed $\dot{U}(\mu_0) \in T_{U(\mu_0)}St(n,p)$. The above equations hold in approximative form for the truncated SVD. If we consider the corresponding parameter-dependent subspaces $[U(\mu_0)]$ in the sense of (5), then the derivative reads

$$\dot{U}(\mu_0)^{\perp} := (I - U(\mu_0)U(\mu_0)^T)\dot{U}(\mu_0) \in T_{[U(\mu_0)]}Gr(n,p). \tag{10}$$

Relationship with the Euclidean Taylor expansion If the snapshot solutions depend real analytically[1] on the operating point μ, then so does the snapshot matrix Y and its singular value decomposition $Y = U\Sigma V^T$, because of the relationship with the symmetric eigenvalue problem $Y^TY = V\Sigma^2V^T$ and the results of [2, 7]. As a consequence, we have a Taylor expansion

$$U(\mu_0 + \mu) = U(\mu_0) + \mu\dot{U}(\mu_0) + \frac{\mu^2}{2}\ddot{U}(\mu_0) + \mathcal{O}(\mu^3) \in St(n,p). \tag{11}$$

[1] Or smoothly in a certain non-pathological sense, see [2, §7].

Note that we leave the Stiefel manifold, if we truncate the Taylor series. The Stiefel geodesic from Algorithm 1, however, matches the Taylor series up to terms of second order while preserving orthonormality, as the next lemma shows.

Lemma 1 *The Stiefel geodesic starting in $U(\mu_0) \in St(n,p)$ with velocity $\mu\dot{U}(\mu_0) \in T_{U(\mu_0)}St(n,p)$ matches the Tayler expansion (11) up to terms of second order,*

$$Exp_U^{St}(\mu\dot{U}(\mu_0)) = U(\mu_0) + \mu\dot{U}(\mu_0) + \mathcal{O}(\mu^2) \in St(n,p).$$

From the Grassmann view point, we have to slightly modify this result.

Lemma 2 *The Grassmann geodesic starting in $[U(\mu_0)] \in Gr(n,p)$ with velocity $\mu\dot{U}(\mu_0)^{\perp} \in T_{[U(\mu_0)]}Gr(n,p)$ matches the Tayler expansion (11) up to the orthogonal part of $\dot{U}(\mu_0)$, see (10), and terms of second order,*

$$Exp_U^{Gr}(\mu\dot{U}(\mu_0)^{\perp}) = [U(\mu_0) + \mu\dot{U}(\mu_0)^{\perp} + \mathcal{O}(\mu^2)] \in Gr(n,p).$$

Comparison with other Taylor-like approaches for local pMOR. The same objective outlined in Sect. 1 is addressed in [6]. In this work, the authors pursue, among others, the approach of a first-order Taylor approximation $U(\mu_0 + \mu) \approx U(\mu_0) + \mu\dot{U}(\mu_0)$. However, $U(\mu_0) + \mu\dot{U}(\mu_0)$ is not in $St(n,p)$. Yet, the departure of $U(\mu_0) + \mu\dot{U}(\mu_0)$ from orthogonality is of the order $\mathcal{O}(\mu^2)$. To see this, recall that $U^T(\mu_0)\dot{U}(\mu_0)$ is skew-symmetric. Hence,

$$(U(\mu_0) + \mu\dot{U}(\mu_0))^T(U(\mu_0) + \mu\dot{U}(\mu_0)) = I + \mu^2\dot{U}(\mu_0)^T\dot{U}(\mu_0).$$

This is a retrospective justification of the approach of [6]. In contrast, the approach proposed here does not rely on a truncated Taylor expansion but computes the Stiefel or Grassmann geodesics explicitly, where the starting velocity is chosen according to the Lemmata 1 and 2, respectively

3 Numerical Illustration

We illustrate the proposed approach by means of an academic example. To this end, we consider the following linear one-dimensional convection-diffusion problem taken from [8]:

$$\partial_t u + \partial_x u = \nu\partial_x^2 u. \tag{12}$$

Here, u is the flow velocity, x is the spatial coordinate and $\nu = 1/Re$ is the reciprocal Reynolds number associated with the viscosity of the fluid. This is a parametric partial differential equation with known closed-form solution [8] given by

$$u(t,x,\nu) = e^{-\alpha(\nu)x}\cos(\beta(\nu)x - t), \tag{13}$$

where $\alpha(\nu) = \frac{1}{4\nu}\left(\sqrt{2+2\sqrt{1+16\nu^2}} - 2\right)$, $\beta(\nu) = \frac{1}{4\nu}\left(\sqrt{-2+2\sqrt{1+16\nu^2}}\right)$.

For experimental purposes, we treat (12) according to the generic procedure outlined in Sect. 1: First, we fix $\nu = \nu_0$. Then we discretize in space with spatial resolution $n \in \mathbb{N}$. Hence, the function $x \mapsto u(t,x,\nu_0)$ is represented as a vector $y(t,\nu_0) \in \mathbb{R}^n$. The next step is a snapshot POD. To this end, we select time instants $t_1,\ldots,t_m$ and compute the snapshot matrix $Y = Y(\nu_0) := (y(t_1,\nu_0),\ldots,y(t_m,\nu_0)) \in \mathbb{R}^{n\times m}$ and the matrix of snapshot derivatives $\dot{Y} = \left(\frac{\partial y(t_1,\nu_0)}{\partial \nu},\ldots,\frac{\partial y(t_m,\nu_0)}{\partial \nu}\right) \in \mathbb{R}^{n\times p}$. Then, we compute $U\Sigma V^T \overset{SVD}{=} Y$ and truncate to obtain $U \in St(n,p), p \le m$. We compute $\dot{U}(\nu_0) = \frac{d}{d\nu}U(\nu_0) \in T_{U(\nu_0)}St(n,p)$ according to (7)–(9).

With $U_0 := U(\nu_0), \dot{U}_0 := \dot{U}(\nu_0)$ at hand, we compute the corresponding Stiefel and Grassmann geodesics according to Algorithms 1 and 2, respectively:

$$U(\nu_0+\nu) \approx Exp^{St}_{U_0}(\nu\dot{U}_0), \quad [U(\nu_0+\nu)] \approx Exp^{Gr}_{U_0}(\nu\dot{U}_0^{\perp}).$$

In this way, we obtain valid Stiefel representatives for ONBs and Grassmann subspace representatives for any small perturbation $\nu \in [0,\delta\nu]$.

A preliminary performance test is conducted with the following settings: The reciprocal Reynolds number is set to $\nu_0 = 0.2$, the spatial resolution is chosen as $n = 1{,}000$, initial snapshots are taken at the time instants of $t_1 = 0.7, t_2 = 0.9, t_3 = 1.1, t_4 = 1.3$.[2] A snapshot SVD leads to a two-dimensional POD subspace represented by $U(\nu_0) \in St(1{,}000,2)$. As an upper bound for the perturbations, we choose $\delta\nu := 0.1\nu_0 = 0.02$.

In order to evaluate the approximation accuracy, we compute the reference POD subspace $U(\nu_0+\nu)$ for $\nu \in [0,\delta\nu]$ by repeating the snapshot POD at each operating point ν based on four snapshots at the same time instants t_1,t_2,t_3,t_4. We consider the following five approximative approaches:

1. $U(\nu_0+\nu) \approx U(\nu_0) \in St(n,p)$ (do not adapt, reuse the subspace at ν_0)
2. $U(\nu_0+\nu) \approx Exp^{St}_U(\nu\dot{U}(\nu_0))$ (parametrize via Stiefel geodesic)
3. $U(\nu_0+\nu) \approx Exp^{Gr}_U(\nu\dot{U}^{\perp}(\nu_0))$ (parametrize via Grassmann geodesic)
4. $U(\nu_0+\nu) \approx svd(Y(\nu_0)+\nu\dot{Y}(\nu))$ (Euclidean Taylor, recomp. of SVD)
5. $U(\nu_0+\nu) \approx qr(U(\nu_0)+\nu\dot{U}(\nu_0))$ (Pseudo Taylor, reorthogonalized.)

The flop count for the approaches 2.–5. is $\mathcal{O}(np^2)$, $\mathcal{O}(np)$, $\mathcal{O}(nm^2)$, $\mathcal{O}(np^2)$. However, the evaluation of the Stiefel and Grassmann geodesics does not require to compute matrix decompositions. In Fig. 2, the gap between the approximate

[2] It is known from theory that the linear convection diffusion problem (12) features exactly two linearly independent POD modes [8]. Taking four snapshots is an intentional oversampling in order to trigger truncation.

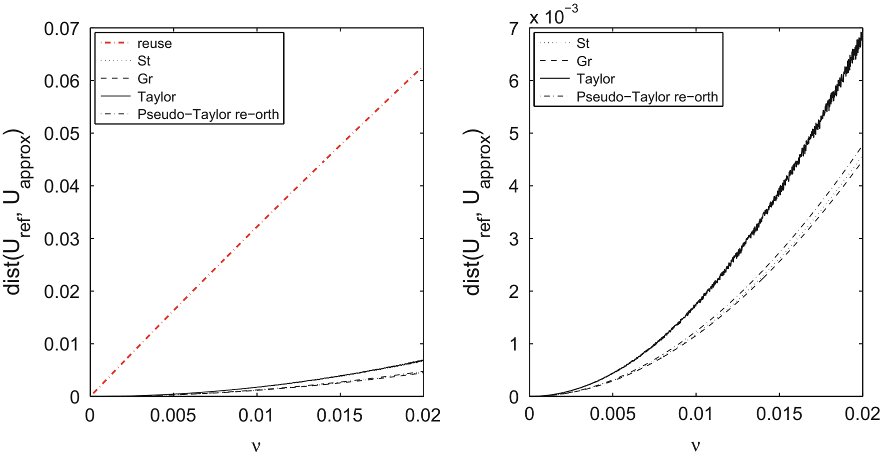


Fig. 2 Plot of the subspace distance $dist(U(\nu_0+\nu), \tilde{U}(\nu_0+\nu))$ vs. $\nu \in [0, \delta\nu]$, where $\tilde{U}(\nu_0+\nu)$ is one of the approximations 1.–5. The geodesic approaches lead to accurate subspaces but come at lower computational costs when compared to those competitors that also lead to orthogonal subspace representatives

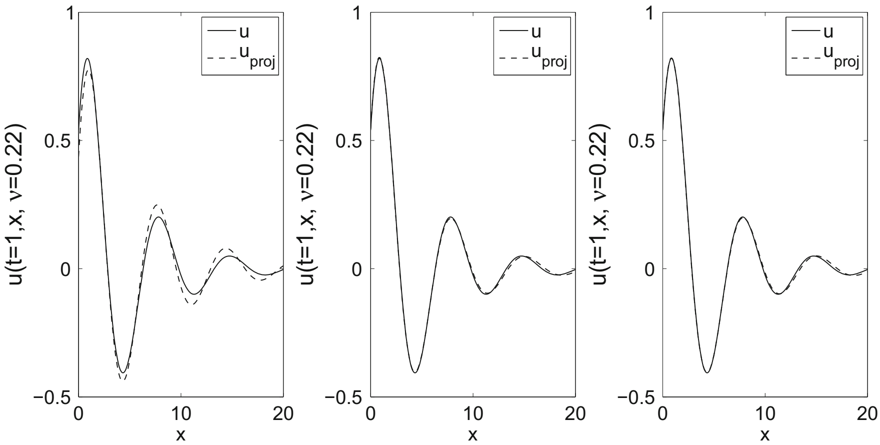


Fig. 3 Reference solution of (12) at $t = 1.0$, $\nu_0 + \delta\nu = 0.22$ (*solid line*) compared to its projection onto the base point subspace $[U(\nu_0)]$ (*left*) and onto the adapted subspaces $[\tilde{U}(\nu_0+\nu)]$ corresponding to the Stiefel geodesic (*middle*) and to the Grassmann geodesic (*right*)

subspaces produced by the approaches 1.–5. and the reference POD subspace is compared in terms of the arc length distance. The arc length distance between two subspaces associated with matrix representatives $U, \tilde{U}$ equals the 2-norm of the vector of principle angles between U and $\tilde{U}$, see [5, §4.3]. As predicted from the theory, the error associated with the first-order approximations 2.–5. grows quadratically in $\nu \in [0, \delta\nu]$, while it grows linearly if we do not adapt the base point subspace. Figure 3 depicts the reference solution $u(t = 1.0, x, \nu = 0.22)$ in comparison with its projection onto the subspaces $[\tilde{U}(\nu_0 + \nu)]$ associated with the approaches 1.,2.,3.

References

1. P.-A. Absil, R. Mahony, R. Sepulchre, *Optimization Algorithms on Matrix Manifolds* (Princeton University Press, Princeton, 2008)
2. D. Alekseevsky, A. Kriegl, P.W. Michor, M. Losik, Choosing roots of polynomials smoothly. Isr. J. Math. **105**(1), 203–233 (1998)
3. P. Benner, S. Gugercin, K. Willcox, A survey of projection-based model reduction methods for parametric dynamical systems. SIAM Rev. **57**(4), 483–531 (2015)
4. A. Bunse-Gerstner, R. Byers, V. Mehrmann, N.K. Nichols, Numerical computation of an analytic singular value decomposition of a matrix valued function. Numer. Math. **60**(1), 1–39 (1991)
5. A. Edelman, T.A. Arias, S.T. Smith, The geometry of algorithms with orthogonality constraints. SIAM J. Matrix Anal. Appl. **20**(2), 303–353 (1999)
6. A. Hay, J.T. Borggaard, D. Pelletier, Local improvements to reduced-order models using sensitivity analysis of the proper orthogonal decomposition. J. Fluid Mech. **629**, 41–72 (2009)
7. T. Kato, *Perturbation Theory for Linear Operators* (Springer, Berlin/Heidelberg, 1995)
8. D.M. Luchtenburg, B.R. Noack, M. Schlegel, An introduction to the POD Galerkin method for fluid flows with analytical examples and MATLAB source codes. Technical report 01/2009, Berlin Institute of Technology, Berlin (2009)

Reduced-Order Multiobjective Optimal Control of Semilinear Parabolic Problems

Laura Iapichino, Stefan Trenz, and Stefan Volkwein

Abstract In this paper a reduced-order strategy is applied to solve a multiobjective optimal control problem governed by semilinear parabolic partial differential equations. These problems often arise in practical applications, where the quality of the system behaviour has to be measured by more than one criterium. The weighted sum method is exploited for defining scalar-valued nonlinear optimal control problems built by introducing additional optimization parameters. The optimal controls corresponding to specific choices of the optimization parameters are efficiently computed by the reduced-basis method. The accuracy is guaranteed by an a-posteriori error estimate.

1 Introduction

In real applications, optimization problems are often described by introducing several objective functions conflicting with each other. This leads to *multiobjective* or *multicriterial* optimization problems; see, e.g., [1]. Finding the optimal control that represents a good compromise is the main issue in these problems. For that reason the concept of Pareto optimal or efficient points is developed. In contrast to scalar-valued optimization problems, the computation of a set of Pareto optimal points is required. Consequently, many scalar-valued constrained optimization problems have to be solved.

When dealing with control functions instead of parameters, a multiobjective optimal control problem (MOCP) needs to be solved. In this paper we apply the

L. Iapichino (✉)
Department of Precision and Microsystems Engineering, Delft University of Technology,
Mekelweg 2, 2628 CD, Delft, The Netherlands
e-mail: l.iapichino@tudelft.nl

S. Trenz • S. Volkwein
Department of Mathematics and Statistics, University of Konstanz, Universitätsstraße 10,
D-78457 Konstanz, Germany
e-mail: stefan.trenz@uni-konstanz.de; stefan.volkwein@uni-konstanz.de

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_37

weighted sum method [1, 13] in order to transform the MOCP into a sequence of scalar optimal control problems and to solve them using well known optimal control techniques [12]. Preliminary results combining reduced-order modeling and multiobjective PDE-constrained optimization are recently derived [4, 9]. The main focus of the present work lies in the extension of [4] to nonlinear, control constrained optimal control problems governed by evolution problems.

The paper is organized as follows. In Sect. 2 the multiobjective optimal control problem is formulated and transformed into a scalar-valued problem that is considered in Sect. 3. The numerical strategy and results are discussed in Sect. 4.

2 Problem Formulation and Pareto Optimality

Let $\Omega \subset \mathbb{R}^{\mathsf{d}}$, $\mathsf{d} \in \{1,2,3\}$, be an open and bounded domain with Lipschitz continuous boundary $\Gamma = \partial\Omega$. For given $T > 0$ we set $Q = (0,T) \times \Omega$ and $\Sigma = (0,T) \times \Gamma$. Then we consider the following class of multiobjective optimal control problems governed by semilinear parabolic equations:

$$\min \mathscr{J}(y,u) = \begin{pmatrix} \mathscr{J}_1(y,u) \\ \mathscr{J}_2(y,u) \\ \mathscr{J}_3(y,u) \end{pmatrix} = \begin{pmatrix} \frac{1}{2}\int_\Omega \left|y(T,\cdot) - y_\Omega\right|^2 \mathrm{d}\boldsymbol{x} \\ \frac{1}{2}\int_0^T \int_\Omega \left|y - y_Q\right|^2 \mathrm{d}\boldsymbol{x}\mathrm{d}t \\ \frac{1}{2}\sum_{i=1}^m \left|u_i - u_i^d\right|^2 \end{pmatrix} \tag{1a}$$

subject to the semilinear parabolic differential problem

$$\begin{aligned} &y_t(t,\boldsymbol{x}) - \Delta y(t,\boldsymbol{x}) + y^3(t,\boldsymbol{x}) = \sum_{i=1}^m u_i b_i(\boldsymbol{x}) + f(t,\boldsymbol{x}) \text{ for } (t,\boldsymbol{x}) \in Q, \\ &\frac{\partial y}{\partial n}(t,\boldsymbol{s}) = 0 \text{ for } (t,\boldsymbol{s}) \in \Sigma, \quad y(0,\boldsymbol{x}) = y_\circ(\boldsymbol{x}) \text{ for } \boldsymbol{x} \in \Omega \end{aligned} \tag{1b}$$

and to the bilateral control constraints

$$u \in \mathscr{U}_{\mathsf{ad}} = \left\{\tilde{u} = (\tilde{u}_1, \ldots, \tilde{u}_m)^\top \in \mathbb{R}^m \,\middle|\, u_i^a \le \tilde{u}_i \le u_i^b \text{ for } 1 \le i \le m\right\}. \tag{1c}$$

In (1a) we assume that $y_\Omega \in L^\infty(\Omega)$, $y_Q \in L^\infty(Q)$, $u^d = (u_1^d, \ldots, u_m^d)^\top \in \mathbb{R}^m$. By '$\top$' we denote the transpose of a vector or matrix. Furthermore, we suppose that $b_1, \ldots, b_m \in L^\infty(\Omega)$, $y_\circ \in L^\infty(\Omega)$. In (1c) let $u_a^i, u_b^i \in \mathbb{R}$ satisfying $u_a^i \le u_b^i$ for $1 \le i \le m$.

Recall that the state equation (1b) has a unique (weak) solution $y = y(u)$ that is bounded for every $u \in \mathscr{U}_{\mathsf{ad}}$; see, e.g., [12]. Hence, we can introduce the reduced objective by $\hat{\mathscr{J}}(u) = \mathscr{J}(y(u), u)$ for $u \in \mathscr{U}_{\mathsf{ad}}$. Instead of (1) we consider now

$$\min \hat{\mathscr{J}}(u) = \begin{pmatrix} \hat{\mathscr{J}}_1(u) \\ \hat{\mathscr{J}}_2(u) \\ \hat{\mathscr{J}}_3(u) \end{pmatrix} \quad \text{subject to} \quad u \in \mathscr{U}_{\mathsf{ad}}. \tag{2}$$

Problem (2) involves the minimization of a vector-valued objective. This is done by using the concepts of order relation and Pareto optimality [1]. In $\mathbb{R}^3$ we make use of the following order relation: For all $z^1, z^2 \in \mathbb{R}^3$ we have

$$z^1 \le z^2 \quad \Leftrightarrow \quad z^2 - z^1 \in \mathbb{R}^3_+ = \big\{ z \in \mathbb{R}^3 \,\big|\, z_i \ge 0 \text{ for } i = 1, 2, 3 \big\}.$$

Definition 1 (Pareto optimal) Let $\mathscr{Z} = \hat{\mathscr{J}}(\mathscr{U}_{\mathsf{ad}}) \subset \mathbb{R}^3$ be the image set of $\mathscr{U}_{\mathsf{ad}}$ under the cost functional $\hat{\mathscr{J}}$. We call a point $\bar{z} \in \mathscr{Z}$ *globally (strictly) efficient* with respect to the order relation $\le$, if there exists no $z \in \mathscr{Z} \setminus \{\bar{z}\}$ with $z \le \bar{z}$. If $\bar{z}$ is efficient and $\bar{u} \in \mathscr{U}_{\mathsf{ad}}$ satisfies $\bar{z} = \hat{\mathscr{J}}(\bar{u})$, we call $\bar{u}$ *(strictly) Pareto optimal*. Let $\bar{u} \in \mathscr{U}_{\mathsf{ad}}$ hold. If there exists a neighborhood $\mathscr{N}(\bar{u}) \subset \mathscr{U}_{\mathsf{ad}}$ of $\bar{u}$ so that $\bar{z} = \hat{\mathscr{J}}(\bar{u})$ is (strictly) efficient for the (local) image set $\hat{\mathscr{J}}(\mathscr{N}(\bar{u})) \subset \mathscr{Z}$, the point $\bar{u}$ is called *locally (strictly) Pareto optimal*. Moreover, $\bar{z}$ is said to be *locally efficient*.

Now, the multiobjective optimal control problem (2) is understood as follows: *Find Pareto optimal points in $\mathscr{U}_{\mathsf{ad}}$ for the vector-valued reduced objective $\hat{\mathscr{J}}$.*

First-order necessary optimality conditions for Pareto optimality are presented in the next theorem which is proved in [1, Theorem 3.21 and Corollary 3.23]. The proof is based on the result of Kuhn-Tucker [6].

Theorem 2 *Suppose that $\bar{u} \in \mathscr{U}_{\mathsf{ad}}$ is Pareto optimal. Then, there exists a parameter vector $\boldsymbol{\mu} = (\bar{\mu}_1, \bar{\mu}_2, \bar{\mu}_3) \in \mathbb{R}^3$ satisfying the Karush-Kuhn-Tucker conditions*

$$\bar{\mu}_i \in [0, 1], \ \sum_{i=1}^{3} \bar{\mu}_i = 1 \text{ and } \sum_{i=1}^{3} \bar{\mu}_i \, \hat{\mathscr{J}}_i'(\bar{u})^\top (u - \bar{u}) \ge 0 \text{ for all } u \in \mathscr{U}_{\mathsf{ad}}. \tag{3}$$

Motivated by Theorem 2, let us choose $0 < \mu_{lb} < 1$ and set

$$\mathscr{D}_{\mathsf{ad}} = \Big\{ \boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3) \in \mathbb{R}^k_+ \,\Big|\, \sum_{i=1}^{3} \mu_i = 1, \ \mu_3 \ge \mu_{lb} \Big\} \subset [0, 1]^3.$$

The condition $\mu_3 \ge \mu_{lb}$ is necessary for the well-posedness of the scalar-valued optimal problem ($\hat{\mathbf{P}}_{\boldsymbol{\mu}}$) introduced below. For any $\boldsymbol{\mu} \in \mathscr{D}_{\mathsf{ad}}$ we define the parameter-dependent, scalar-valued objective as $\hat{J}(u; \boldsymbol{\mu}) = \boldsymbol{\mu}^\top \hat{\mathscr{J}}(u)$ for $u \in \mathscr{U}_{\mathsf{ad}}$. Then, (3) are the first-order necessary optimality conditions for a local solution $\bar{u} = \bar{u}(\boldsymbol{\mu})$ to the parameter-dependent optimization problem

$$\min \hat{J}(u;\boldsymbol{\mu}) \quad \text{subject to} \quad u \in \mathscr{U}_{\mathsf{ad}} \tag{$\hat{\mathbf{P}}_{\boldsymbol{\mu}}$}$$

for the parameter $\boldsymbol{\mu} = \bar{\boldsymbol{\mu}}$. In the *weighted sum method* Pareto optimal points are computed by solving ($\hat{\mathbf{P}}_{\boldsymbol{\mu}}$) for various $\boldsymbol{\mu} \in \mathscr{D}_{\mathsf{ad}}$; see [13] and [1, Chapter 3].

Remark 3 To solve ($\hat{\mathbf{P}}_{\boldsymbol{\mu}}$) we apply a globalized Newton-CG method [7]. ◇

3 The Scalar-Valued Optimal Control Problem

Suppose that $\bar{u} = u(\boldsymbol{\mu}) \in \mathscr{U}_{\mathsf{ad}}$ is an optimal solution to ($\hat{\mathbf{P}}_{\boldsymbol{\mu}}$) for given $\boldsymbol{\mu} \in \mathscr{D}_{\mathsf{ad}}$. Let $\bar{y} = y(\bar{u};\boldsymbol{\mu})$ denote the associated optimal state satisfying (1b) for $u = \bar{u}$. Following [12], first-order necessary optimality conditions for ($\hat{\mathbf{P}}_{\boldsymbol{\mu}}$) ensure the existence of a unique adjoint $\bar{p} = p(\bar{u};\boldsymbol{\mu})$ solving

$$\begin{aligned} &-p_t(t,\boldsymbol{x}) - \Delta p(t,\boldsymbol{x}) + 3y^2(t,\boldsymbol{x})p(t,\boldsymbol{x}) = \mu_1\big(y_Q(t,\boldsymbol{x}) - \bar{y}(t,\boldsymbol{x})\big) \text{ in } Q,\\ &\frac{\partial p}{\partial n}(t,\boldsymbol{s}) = 0 \text{ on } \Sigma, \quad p(T,\boldsymbol{x}) = \mu_2\big(y_\Omega(\boldsymbol{x}) - \bar{y}(T,\boldsymbol{x})\big) \text{ in } \Omega \end{aligned} \tag{4}$$

with $y = \bar{y}$. Moreover, for any $\boldsymbol{\mu} \in \mathscr{D}_{\mathsf{ad}}$ the gradient of the reduced cost functional $\hat{J}(\cdot\,;\boldsymbol{\mu})$ at a given $u \in \mathscr{U}_{\mathsf{ad}}$ is given by

$$\hat{J}'(u;\boldsymbol{\mu}) = \left(\mu_3\big(u_i - u_i^d\big) - \int_0^T\int_\Omega p(t,\boldsymbol{x})b_i(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\mathrm{d}t\right)_{1\le i\le m},$$

where p solves (4) and y is the solution to (1b).

Since ($\hat{\mathbf{P}}_{\boldsymbol{\mu}}$) is a non-convex problem, we make use of the hessian $\hat{J}''(u;\boldsymbol{\mu}) \in \mathbb{R}^{m\times m}$ in order to ensure sufficient optimality conditions. Let $\bar{u} = \bar{u}(\boldsymbol{\mu})$ be locally optimal for ($\hat{\mathbf{P}}_{\boldsymbol{\mu}}$) satisfying the second-order sufficient optimality condition

$$\hat{J}''(\bar{u};\boldsymbol{\mu})(u,u) \ge \kappa\,|u|_2^2 \quad \text{for all } u \in \mathbb{R}^m$$

with a $\boldsymbol{\mu}$-independent lower bound $\kappa > 0$ for the smallest eigenvalue of the hessian. Then, for any $\tilde{\kappa} \in (0,\kappa)$ there exists a radius $\tilde{\rho} = \rho(\tilde{\kappa}) > 0$ such that

$$\hat{J}''(\tilde{u};\boldsymbol{\mu})(u,u) \ge \tilde{\kappa}\,|u|_2^2 \quad \text{for all } \tilde{u} \text{ with } |\tilde{u} - \bar{u}|_2 \le \tilde{\rho}. \tag{5}$$

If $\tilde{u} \in \mathscr{U}_{\mathsf{ad}}$ satisfies (5) we can estimate the error between $\bar{u}$ and $\tilde{u}$ as [5]

$$|\bar{u} - \tilde{u}|_2 \le \frac{1}{\tilde{\kappa}}\,|\zeta|_2, \tag{6}$$

where $\zeta = \zeta(\tilde{u}) \in \mathbb{R}^m$ is given by

$$\zeta_i := \begin{cases} \Big[\mu_3\big(u_i - u_i^d\big) - \int_0^T \int_\Omega \tilde{p}(t,\boldsymbol{x}) b_i(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}\mathrm{d}t\Big]_-, & \text{if } \tilde{u}_i = u_i^a, \\ -\mu_3\big(u_i - u_i^d\big) + \int_0^T \int_\Omega \tilde{p}(t,\boldsymbol{x}) b_i(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}\mathrm{d}t, & \text{if } u_i^a < \tilde{u}_i < u_i^b, \\ -\Big[\mu_3\big(u_i - u_i^d\big) - \int_0^T \int_\Omega \tilde{p}(t,\boldsymbol{x}) b_i(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}\mathrm{d}t\Big]_+, & \text{if } \tilde{u}_i = u_i^b, \end{cases} \tag{7}$$

$\tilde{p}$ solves (4) with $y = \tilde{y}$ and $\tilde{y}$ solves (1b) with $u = \tilde{u}$. In (7) we denote by $[s]_- = -\min(0, s)$ and $[s]_+ = \max(0, s)$ the negative and positive part function, respectively. Hence, given a (suboptimal) $\tilde{u} \in \mathscr{U}_{\mathsf{ad}}$, the error $\bar{u} - \tilde{u}$ can be estimated by the right-hand side in (5) provided the lower bound $\tilde{\kappa}$ for the symmetric matrix $\hat{J}''(\tilde{u}; \boldsymbol{\mu})$ is known. We proceed as follows: From $\hat{J}(\tilde{u}; \boldsymbol{\mu}) = \boldsymbol{\mu}^\top \hat{\mathscr{J}}(\tilde{u})$ we find

$$\hat{J}''(\tilde{u}; \boldsymbol{\mu}) = \mu_1 \hat{\mathscr{J}}_1''(\tilde{u}) + \mu_2 \hat{\mathscr{J}}_2''(\tilde{u}) + \mu_3 \hat{\mathscr{J}}_3''(\tilde{u}) \quad \text{for } \boldsymbol{\mu} \in \mathscr{D}_{\mathsf{ad}}.$$

It follows from Theorem of Courant-Fischer [10, Corollary 4.7] that a lower bound $\lambda_{\min}^{\mathrm{LB}}$ for the smallest eigenvalue $\lambda_{\min}$ of $\hat{J}''(\tilde{u}; \boldsymbol{\mu})$ is given by

$$\lambda_{\min}\big(\hat{J}''(\tilde{u}; \boldsymbol{\mu})\big) \geq \sum_{i=1}^{3} \mu_i \lambda_{\min}\big(\hat{\mathscr{J}}_i''(\tilde{u})\big) =: \lambda_{\min}{}^{\mathrm{LB}}(\tilde{u}; \boldsymbol{\mu}), \tag{8}$$

where $\lambda_{\min}(A)$ denotes the smallest eigenvalue of a symmetric matrix A.

4 Numerical Solution Strategy

To solve (1) we apply the weighted sum method. Thus, the set of Pareto optimal points is approximated by solutions $\bar{u}(\boldsymbol{\mu})$ to ($\hat{\mathbf{P}}_{\boldsymbol{\mu}}$) for various parameters $\boldsymbol{\mu} \in \mathscr{D}_{\mathsf{ad}}$. Consequently, many constrained nonlinear optimization problems have to be solved numerically, which is computationally expensive. For this reason, model-order reduction (MOR) is applied to reduce significantly the required computational resources. Our MOR approach is based on a Galerkin-type approximation to ($\hat{\mathbf{P}}_{\boldsymbol{\mu}}$) using MOR basis functions, where for certain weighting parameters $\boldsymbol{\mu} \in \mathscr{D}_{\mathsf{ad}}$ the MOR basis functions contain information from optimal states $\bar{y}(\boldsymbol{\mu})$ and adjoints $\bar{p}(\boldsymbol{\mu})$ associated with optimal controls $\bar{u}(\boldsymbol{\mu})$. The MOR basis functions are determined in an *offline phase*. In the *online phase* the weighted sum method is applied, where numerical solutions to ($\hat{\mathbf{P}}_{\boldsymbol{\mu}}$) are computed rapidly by a MOR Galerkin discretization [11].

Offline phase I: eigenvalue computation on control grid Let us choose a discrete (regular) control grid $\Xi_{\text{grid}} = \{u^k\}_{k=1}^K$ in the set $\mathscr{U}_{\mathsf{ad}}$ of admissible controls. In an *offline* phase we compute and store the $\boldsymbol{\mu}$-independent smallest eigenvalues

$\lambda_{\min}(\hat{\mathscr{J}}_i''(u^k))$ at any grid node $u^k \in \varXi_{\mathrm{grid}}$ for $i = 1$ and 2. Since $\hat{\mathscr{J}}_3''(u^k)$ is the identity, we have $\lambda_{\min}(\hat{\mathscr{J}}_3''(u^k)) = 1$. Now, (8) yields a numerically efficient computation of the approximative lower bound $\lambda_{\mathrm{app}}^{\mathrm{LB}}(\tilde{u};\boldsymbol{\mu})$ in the *online* phase at any (suboptimal) control $\tilde{u} \in \mathscr{U}_{\mathsf{ad}}$ by convex combination of the stored smallest eigenvalues $\lambda_{\min}(\hat{\mathscr{J}}_i''(u^k))$ for $k = 1,\ldots,K$:

$$\lambda_{\min}\big(\hat{\mathscr{J}}_i''(\tilde{u})\big) \approx \lambda_{\mathrm{app}}^i(\tilde{u}) := \sum_{k=1}^{K} \omega_k\, \lambda_{\min}\big(\hat{\mathscr{J}}_i''(u^k)\big). \tag{9}$$

In (6) we utilize $\lambda_{\mathrm{app}}^i(\tilde{u})$ instead of $\tilde{\kappa}$.

Remark 4 The computation of the $\lambda_{\min}(\hat{\mathscr{J}}_i''(u^k))$ can be realized in parallel computing with respect to $k \in \{1,\ldots,K\}$. ◇

Offline phase II: MOR basis computation Estimate (6) can be suitably used as ingredient to apply a MOR strategy for the solution of nonlinear multiobjective problems. In order to use a MOR technique for its solution, we propose to use the POD-greedy algorithm based on [3] and [2, 8]. As an input the POD-greedy algorithm requires a discrete parameter training set $\mathsf{S}_{\mathrm{train}} \subset \mathscr{D}_{\mathsf{ad}}$, as well as he smallest eigenvalues $\lambda_{\min}(\hat{\mathscr{J}}_i''(u^k))$ for $i = 1,2,3$ on the control grid $\varXi_{\mathrm{grid}}$ and the corresponding precomputed grid node data $\mathsf{D}_{\mathrm{grid}}$, both needed for the smallest eigenvalue approximation in the a-posteriori error estimation.

Online phase: multiobjective optimal control As regards the original multiobjective problem, we are interested in the solution of the parametric optimal control problem for a large number of parameter values, since we want to identify a set of optimal control solutions that does not a-priorily penalize any cost functional. In other words, we are interested in identifying the Pareto optimal front of the multiobjective problem, that consists in a large set of cost functionals evaluation corresponding to the solution of a large number of optimal control problems (obtained in correspondence of several parameter values, at the randomly chosen). In order to identify the Pareto front we need to evaluate several times the parametric optimal control problem. For this purpose, the proposed model order reduction strategy can be efficiently reduce the required computational times.

5 Numerical Example

We consider (1) with spatial domain $\varOmega = (0,1) \times (0,1) \subset \mathbb{R}^2$, final time $T = 1$, desired states $y_\varOmega = 0$, $y_Q(t,\boldsymbol{x}) = 100t\cos(2\pi x_1)\cos(2\pi x_2)$, initial condition $y_\circ(\boldsymbol{x}) = 0$ and inhomogeneity $f(t,\boldsymbol{x}) = 10tx_1$. Furthermore, for $m = 4$ each shape function $b_i(\boldsymbol{x}), i = 1,\ldots,4$, has the support in a quarter of the domain $\varOmega$ and $u^d = (0.5,-4,-0.5,4)^\top \in \mathbb{R}^4$. The high fidelity spatial approximations of the problem solution, used for the basis computations and the error comparisons is computed by a finite element (FE) model with a Newton method that uses $\mathbb{P}_1$ elements, it has 729 degrees of freedom in 1352 elements. For the temporal discretization the implicit Euler method is applied with equidistant step size $\Delta t = 0.01$ steps. Figure 1 shows optimal states $\bar{y} = y(\bar{u}; \boldsymbol{\mu})$ for solutions $\bar{u}(\boldsymbol{\mu})$ to $(\hat{\mathbf{P}}_\mu)$ corresponding to two values of the parameter $\boldsymbol{\mu}$. In the left plot of Fig. 2 we plot error comparisons obtained by solving $(\hat{\mathbf{P}}_\mu)$.

In particular, we compare the errors between the optimal controls computed by the model order reduction (MOR) method and the one by the FE method. We show its minimum, maximum and average values (over a range of 1000 parameter values) by varying the number of basis functions used in the MOR scheme. In the right plot of Fig. 2 we present the Pareto front obtained by solving $(\hat{\mathbf{P}}_\mu)$ with the MOR technique for different parameter values $\boldsymbol{\mu}$. In order to show the correctness of the Pareto front, we also include the values of the cost functionals $\hat{\mathscr{J}}_1(u)$, $\hat{\mathscr{J}}_2(u)$ and $\hat{\mathscr{J}}_3(u)$ obtained for 1000 control values randomly chosen as follows: $u_1 \in [-3, 3], u_2 \in [-8, -1], u_3 \in [-5, -2], u_4 \in [-1, 6]$ (not optimal controls). In

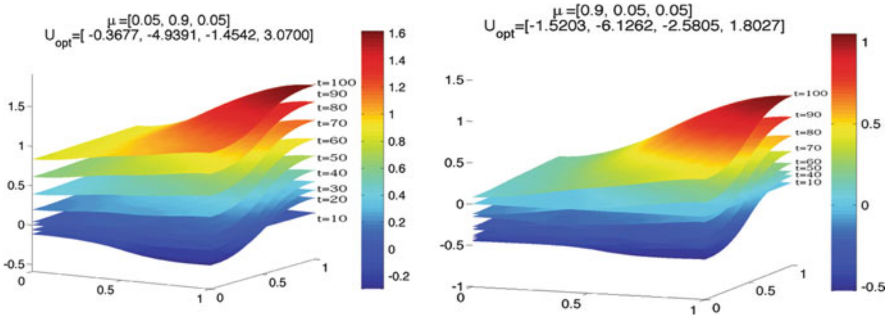


Fig. 1 Optimal states $\bar{y} = y(\bar{u}; \boldsymbol{\mu})$ for solutions $\bar{u}(\boldsymbol{\mu})$ to $(\hat{\mathbf{P}}_\mu)$ for parameter values $\boldsymbol{\mu} = (0.05, 0.9, 0.05)$ (*left*) and $\boldsymbol{\mu} = (0.9, 0.05, 0.05)$ (*right*)

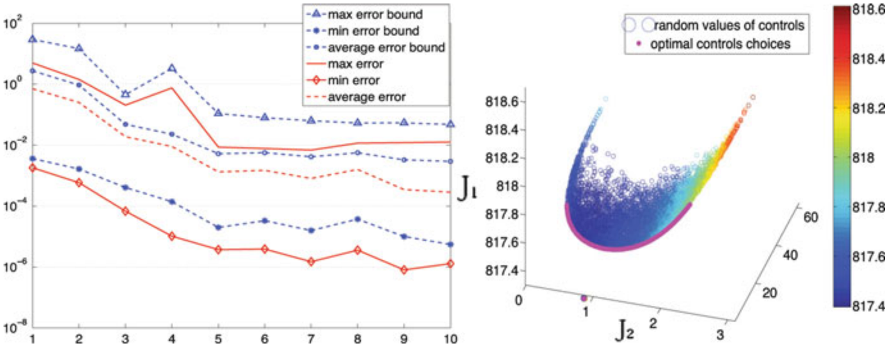


Fig. 2 Error bound and error between the MOR and FE solutions by varying the number of bases: minimum, maximum and average values over a set of 100 random parameter values (*left*); Pareto front and cost functionals values corresponding to admissible control points (*right*)

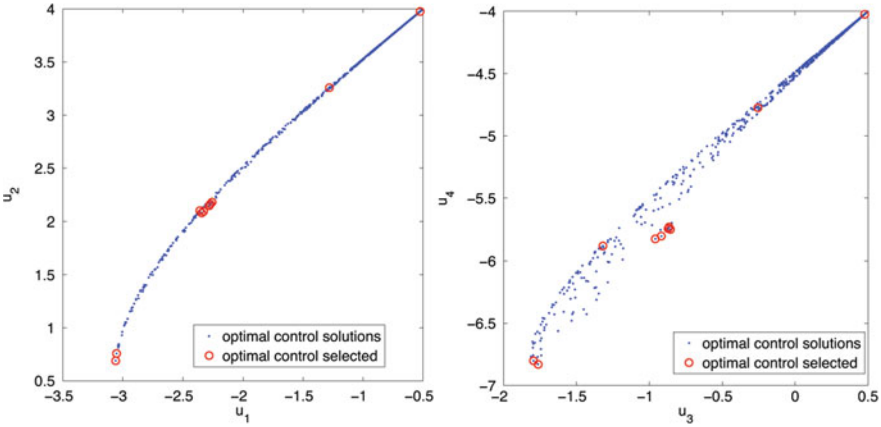


Fig. 3 Optimal controls $\bar{u}(\boldsymbol{\mu})$ corresponding to 1000 randomly chosen parameter values (*left*) and to the parameter values selected for the bases computations during the POD-greedy algorithm (*right*)

Fig. 3 we show the optimal controls $\bar{u}(\boldsymbol{\mu})$ corresponding to 1000 randomly chosen parameter values and to the parameter values selected for the bases computations during the POD-greedy algorithm. Regarding the computational performances the online evaluation time required for solving problem ($\hat{\mathbf{P}}_{\mu}$) for each parameter value by using 10 basis functions is about 0.7 s; while the evaluation of the FE solution requires about 7.3 s. We conclude that the gained speedup allows a much faster optimal solution evaluation and an efficient identification of the Pareto front, for which several repeated solutions have to be computed.

References

1. M. Ehrgott, *Multicriteria Optimization* (Springer, Berlin, 2005)
2. M. Grepl, Certified reduced basis methods for nonaffine linear time-varying and nonlinear parabolic partial differential equations. Math. Models Methods Appl. Sci. **22**(3), 1150015 (2012)
3. B. Haasdonk, M. Ohlberger, Reduced basis method for finite volume approximations of parametrized linear evolution equations. Math. Model. Numer. Anal. **42**, 277–302 (2008)
4. L. Iapichino, S. Ulbrich, S. Volkwein, Multiobjective PDE-constrained optimization using the reduced-basis method. Submitted (2015). Available at http://nbn-resolving.de/urn:nbn:de:bsz:352-2501909
5. E. Kammann, F. Tröltzsch, S. Volkwein, A-posteriori error estimation for semilinear parabolic optimal control problems with application to model reduction by POD. Math. Model. Numer. Anal. **47**, 555–581 (2013)
6. H. Kuhn, A. Tucker, Nonlinear programming, in *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, ed. by J. Newman (University of California Press, Berkeley, 1951), pp. 481–492

7. J. Nocedal, S.J. Wright, *Numerical Optimization*. Springer Series in Operation Research, 2nd edn. (Springer, New York, 2006)
8. A.T. Patera, G. Rozza, *Reduced Basis Approximation and a Posteriori Error Estimation for Parametrized Partial Differential Equations* (MIT, Cambridge, 2007)
9. S. Peitz, M. Dellnitz, Multiobjective optimization of the flow around a cylinder using model order reduction, in *86th Annual Meeting of the International Association of Applied Mathematics and Mechanics (GAMM)*, Lecce, vol. 15, Issue 1 (2015), pp. 613–614
10. G.W. Stuart, J. Sun, *Matrix Perturbation Theory. Computer Science and Scientific Computing* (Academic Press, Boston/San Diego/New York, 1990)
11. E. Sachs, S. Volkwein, POD Galerkin approximations in PDE-constrained optimization. GAMM-Mitt. **33**, 194–208 (2010)
12. F. Tröltzsch, *Optimal Control of Partial Differential Equations: Theory, Methods and Applications*. Graduate Studies in Mathematics, vol. 112 (American Mathematical Society, Providence, 2010)
13. L. Zadeh, Optimality and non-scalar-valued performance criteria. IEEE Trans. Autom. Control **8**, 59–60 (1963)

Part VI
Problems with Singularities

Coupling Fluid-Structure Interaction with Phase-Field Fracture: Modeling and a Numerical Example

Thomas Wick

Abstract In this work, a framework for coupling arbitrary Lagrangian-Eulerian fluid-structure interaction with phase-field fracture is suggested. The key idea is based on applying the weak form of phase-field fracture, including a crack irreversibility constraint, to the nonlinear coupled system of Navier-Stokes and elasticity. The resulting setting is formulated via variational-monolithic coupling and has four unknowns: velocities, displacements, pressure, and a phase-field variable. The inequality constraint is imposed through penalization using an augmented Lagrangian algorithm. The nonlinear problem is solved with Newton's method. The framework is tested in terms of a numerical example in which computational stability is demonstrated by evaluating goal functionals on different spatial meshes.

1 Introduction

Both fluid-structure interaction (FSI) and fracture propagation are current but challenging topics with numerous applications in applied mathematics and engineering. In this work, we want to bring both frameworks together. The idea is to employ the nowadays standard arbitrary Lagrangian-Eulerian (ALE) technique [8, 13] for coupling the isothermal, incompressible Navier-Stokes equations with the geometrically nonlinear Saint Venant-Kirchhoff model. The resulting formulation using variational-monolithic coupling is outlined in [12, 19]. Here, three unknowns are sought: velocities, pressure and displacements. On the other hand, brittle fracture propagation using variational techniques has attracted attention since the pioneering work in [4, 9]. Specifically, we consider a pressurized fracture as it has been formulated in [16]. Here, the crack irreversibility constraint has been imposed through penalization. In phase-field fracture, two unknowns are sought: displacements and a phase-field function that determines the crack location. A first (minor) novelty of this paper is an augmented Lagrangian penalization for a fully-coupled phase-field fracture framework. This is in contrast to [18] in which the

T. Wick (✉)
RICAM, Austrian Academy of Sciences, Altenberger Str. 69, 4040 Linz, Austria
e-mail: thomas.wick@ricam.oeaw.ac.at

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_38

displacement phase-field system has been solved in a partitioned fashion. Such a technique has been employed due to the fact that the underlying energy functional is non-convex in both variables simultaneously [3, 5], which causes serious challenges in the numerical solution. Recently, a robust (but heuristic) methodology of a quasi fully-coupled approach has been proposed in [10] where the phase-field variable has been time-lagged in the displacement equation.

Since the solid part of FSI is based on elastodynamics, we accentuate the work of [2, 6, 14] who extended variational quasi-static brittle fracture to dynamic brittle fracture taking into account the solid acceleration term. Collecting all these different pieces allows us to apply the phase-field fracture technique to the solid part of the FSI problem. More specifically, the phase-field part is re-written (similarly to the flow problem in FSI) in ALE coordinates. The resulting formulation is consequently prescribed in a fixed, but arbitrary, reference domain and all coupling conditions are satisfied in a variational exact fashion on the continuous level. The numerical discretization is then straightforward as the Rothe method (first time, then space) can be applied on the resulting semilinear form. The nonlinear coupled problem is solved with Newton's method. The outline of this paper is as follows: In Sect. 2, the equations are gathered; followed by brief hints in Sect. 3 on the discretization and numerical solution. The framework is tested in Sect. 4 with a prototype example in which computational stability for certain goal functional values is shown.

2 Notation, Spaces, Equations

We denote by $\Omega := \Omega(t) \subset \mathbb{R}^d$, $d = 2$, the domain of the FSI phase-field fracture problem; see Fig. 1. This domain consists of three time-dependent subdomains $\Omega_f(t)$, $\Omega_s(t)$ and $\mathscr{C}(t)$. We assume $\mathscr{C}(t) \subset\subset \Omega_s(t)$. The FSI-interface between $\Omega_f(t)$ and $\Omega_s(t)$ is denoted by $\Gamma_i(t) = \partial\Omega_f(t) \cap \partial\Omega_s(t)$. The initial (or later reference) domains are denoted by $\widehat{\Omega}, \widehat{\Omega}_f$ and $\widehat{\Omega}_s$, respectively, with the interface $\widehat{\Gamma}_i =$

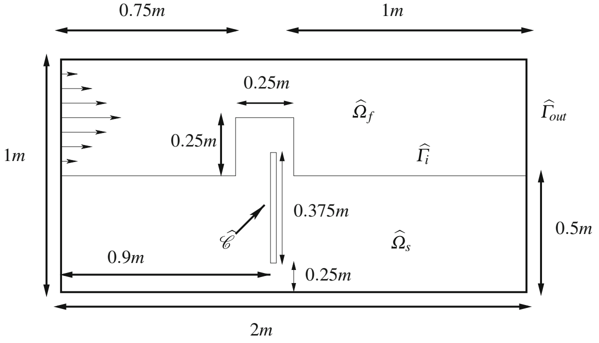


Fig. 1 Configuration and notation

$\partial\widehat{\Omega}_f \cap \partial\widehat{\Omega}_s$. Furthermore, we denote the outer boundary by $\partial\widehat{\Omega} = \widehat{\Gamma} = \widehat{\Gamma}_D \cup \widehat{\Gamma}_{out}$ where $\widehat{\Gamma}_D$ and $\widehat{\Gamma}_{out}$ denote Dirichlet and outflow Neumann boundaries, respectively. Specifically, $\widehat{\Gamma}_{out}$ is motivated by [11]. We denote the L^2 scalar product with $(\cdot,\cdot)$ as frequently used in the literature. Finally, let T the end time value. For the function spaces, we set:

$$\begin{aligned}
\hat{L}_f &:= L^2(\widehat{\Omega}_f), \quad \hat{L}_s := L^2(\widehat{\Omega}_s), \quad \hat{L}_f^0 := L^2(\widehat{\Omega}_f)/\mathbb{R}, \quad \hat{L}_s^0 := L^2(\widehat{\Omega}_s)/\mathbb{R},\\
\hat{V}_f^0 &:= H_0^1(\widehat{\Omega}_f), \quad \hat{V}_{f,\hat{v}}^0 := \{\hat{v}_f \in H_0^1(\widehat{\Omega}_f) : \hat{v}_f = \hat{v}_s \text{ on } \widehat{\Gamma}_i\},\\
\hat{V}_s^0 &:= H_0^1(\widehat{\Omega}_s), \quad \hat{V}_{f,\hat{u}}^0 := \{\hat{u}_f \in H_0^1(\widehat{\Omega}_f) : \hat{u}_f = \hat{u}_s \text{ on } \widehat{\Gamma}_i\},\\
\hat{V}_{f,\hat{u},\widehat{\Gamma}_i}^0 &:= \{\hat{\psi}_f \in H_0^1(\widehat{\Omega}_f) : \hat{\psi}_f = \hat{\psi}_s \text{ on } \widehat{\Gamma}_i \subset \partial X\},\\
W &:= \{\varphi \in H^1(\Omega_s \cup \mathscr{C}) : \partial_t\varphi \leq 0 \text{ a.e. on } \Omega_s \cup \mathscr{C}\}.
\end{aligned}$$

2.1 Variational-Monolithic ALE Fluid-Structure Interaction

First, we define the ALE transformation:

Definition 1 The ALE mapping is defined in terms of the vector-valued fluid mesh displacement $\hat{u}_f$ (obtained by solving a mesh motion problem) such that

$$\hat{\mathscr{A}}(\hat{x},t) : \widehat{\Omega}_f \times I \to \Omega_f, \quad \text{with } \hat{\mathscr{A}}(\hat{x},t) = \hat{x} + \hat{u}_f(\hat{x},t), \tag{1}$$

which is specified through the deformation gradient and its determinant

$$\widehat{F} := \widehat{\nabla}\hat{\mathscr{A}} = \hat{I} + \widehat{\nabla}\hat{u}_f, \quad \hat{J} := \det(\widehat{F}). \tag{2}$$

Here, $\hat{I}$ denotes the identity matrix. The mesh velocity is defined by $\hat{w} := \partial_t\hat{\mathscr{A}}$ and is numerically realized as $\hat{w} = k^{-1}(\hat{u}_f - \hat{u}_f^{n-1})$, where $\hat{u}_f$ is the current displacement solution and $\hat{u}_f^{n-1}$ the previous time step solution, and $k := t^n - t^{n-1}$ being the time step size. The key quantity to measure the fluid mesh regularity is $\hat{J}$.

The weak form of a variational-monolithic FSI model reads, e.g., [19]:

Proposition 2 *Find vector-valued velocities, vector-valued displacements and a scalar-valued fluid pressure, i.e.,* $\{\hat{v}_f, \hat{v}_s, \hat{u}_f, \hat{u}_s, \hat{p}_f\} \in \{\hat{v}_f^D + \hat{V}^0_{f,\hat{v}}\} \times \hat{L}_s \times \{\hat{u}_f^D + \hat{V}^0_{f,\hat{u}}\} \times \{\hat{u}_s^D + \hat{V}^0_s\} \times \hat{L}^0_f$*, such that* $\hat{v}_f(0) = \hat{v}^0_f$*,* $\hat{v}_s(0) = \hat{v}^0_s$*,* $\hat{u}_f(0) = \hat{u}^0_f$*, and* $\hat{u}_s(0) = \hat{u}^0_s$ *are satisfied, and for almost all time steps holds:*

$$\begin{cases} (\hat{J}\hat{\rho}_f \partial_t \hat{v}_f, \hat{\psi}^v)_{\widehat{\Omega}_f} + (\hat{\rho}_f \hat{J}(\widehat{F}^{-1}(\hat{v}_f - \hat{w}) \cdot \widehat{\nabla})\hat{v}_f, \hat{\psi}^v)_{\widehat{\Omega}_f} + (\hat{J}\hat{\sigma}_f \widehat{F}^{-T}, \hat{\nabla}\hat{\psi}^v)_{\widehat{\Omega}_f} \\ + \langle \hat{\rho}_f \nu_f \hat{J}(\widehat{F}^{-T}\hat{\nabla}\hat{v}_f^T \hat{n}_f)\widehat{F}^{-T}, \hat{\psi}^v \rangle_{\widehat{\Gamma}_{out}} = 0 \quad \forall \hat{\psi}^v \in \hat{V}^0_{f,\widehat{\Gamma}_i}, \end{cases}$$

$$\begin{cases} (\hat{\rho}_s \partial_t \hat{v}_s, \hat{\psi}^v)_{\widehat{\Omega}_s} + (\widehat{F}\widehat{\Sigma}, \hat{\nabla}\hat{\psi}^v)_{\widehat{\Omega}_s} \quad = 0 \quad \forall \hat{\psi}^v \in \hat{V}^0_s, \end{cases}$$

$$\begin{cases} (\hat{\sigma}_{mesh}, \hat{\nabla}\hat{\psi}^u)_{\widehat{\Omega}_f} \quad = 0 \quad \forall \hat{\psi}^u \in \hat{V}^0_{f,\hat{u},\widehat{\Gamma}_i}, \end{cases}$$

$$\begin{cases} \hat{\rho}_s (\partial_t \hat{u}_s - \hat{v}_s, \hat{\psi}^u)_{\widehat{\Omega}_s} \quad = 0 \quad \forall \hat{\psi}^u \in \hat{L}_s, \end{cases}$$

$$\begin{cases} (\widehat{div}\,(\hat{J}\widehat{F}^{-1}\hat{v}_f), \hat{\psi}^p)_{\widehat{\Omega}_f} \quad = 0 \quad \forall \hat{\psi}^p \in \hat{L}^0_f. \end{cases}$$

The stress tensors for fluid, solid and mesh motion read:

$$\hat{\sigma}_f = -\hat{p}_f \hat{I} + 2\hat{\rho}_f \nu_f (\hat{\nabla}\hat{v}_f \widehat{F}^{-1} + \widehat{F}^{-T}\hat{\nabla}\hat{v}_f^T),$$

$$\widehat{\Sigma} = 2\mu_s \widehat{E} + \lambda_s tr\widehat{E}\hat{I}, \quad \text{with the strain } \widehat{E} = \frac{1}{2}(\widehat{F}^T\widehat{F} - \hat{I}),$$

$$\hat{\sigma}_{mesh} = \hat{J}^{-1}\alpha_u \hat{\nabla}\hat{u}_f,$$

with the densities $\hat{\rho}_f$ *and* $\hat{\rho}_s$*, fluid's viscosity* ν_f*. The solid parameters are given by the Lamé parameters* μ_s*,* λ_s *and the normal vector is* $\hat{n}_f$*. Finally,* $\hat{J}^{-1}\alpha_u > 0$ *is used to control the fluid mesh motion.*

2.2 Variational Phase-Field for Dynamic Pressurized-Fractures

In phase-field-based fracture propagation, the unknown solution variables are displacements $u : \Omega_s \cup \overline{\mathscr{C}} \to \mathbb{R}^2$ and a smoothed indicator phase-field function $\varphi : \Omega_s \cup \overline{\mathscr{C}} \to [0, 1]$. Here $\varphi = 0$ denotes the crack region and $\varphi = 1$ characterizes the unbroken material. The intermediate values constitute a smooth transition zone dependent on a regularization parameter ε. The physics of the underlying problem ask to enforce a crack irreversibility condition (the crack can never heal) that is an inequality condition in time: $\partial_t \varphi \leq 0$. Consequently, modeling of fracture evolution problems leads to a variational inequality system, that is, due to this constraint, quasi-stationary or time-dependent. Our system of equations applies to cracks in elasticity and pressurized fractures. In the latter one, a Neumann condition acts on the crack surface [16]. Using Gauss' divergence theorem the pressure can be formulated as a domain term: $p_F : \Omega_s \cup \mathscr{C} \to \mathbb{R}$. We now extend this quasi-static pressurized fracture model, while additionally including the solid acceleration term:

Proposition 3 *Let $p_F \in H^1(\Omega_s \cup \mathscr{C})$ and $\tilde{\varphi}$ (see [10]) be given. Find $(u, \varphi) \in \{u_D + H_0^1(\Omega_s \cup \mathscr{C})\} \times W$ for almost all times $t \in (0, T]$ such that*

$$\Big\{(\rho_s \partial_t^2 u, \psi^u) + \Big(\big((1-\kappa)\tilde{\varphi}^2 + \kappa\big)\ \Sigma(u), \nabla\psi^u\Big) + (\tilde{\varphi}^2 p_F, \nabla \cdot \psi^u) = 0\ \forall \psi^u \in V,$$

$$\begin{cases} (1-\kappa)(\varphi\ \Sigma(u) : e(u), \psi^\varphi - \varphi) + 2(\varphi\ p_F\ \nabla \cdot u, \psi^\varphi - \varphi) \\ +G_c\Big(-\frac{1}{\varepsilon}(1-\varphi, \psi^\varphi - \varphi) + \varepsilon(\nabla\varphi, \nabla(\psi^\varphi - \varphi))\Big) \geq 0\ \forall \psi^\varphi \in W \cap L^\infty(\Omega_s \cup \mathscr{C}). \end{cases}$$

In Proposition 3, κ is a positive regularization parameter for the elastic energy, with $\kappa = o(\varepsilon)$, and G_c is the critical energy release rate. Furthermore, Σ is the linearized version of $\widehat{\Sigma}$ with $E = \frac{1}{2}(\nabla u + \nabla u^T)$.

2.3 The Final System

We collect all pieces and perform two additional steps in Proposition 3:

- transforming from $\Omega_s \cup \mathscr{C}$ to $\widehat{\Omega}_s \cup \widehat{\mathscr{C}}$ and re-defining the space W in terms of $\hat{W}$, respectively;
- introducing the augmented Lagrangian penalization strategy to treat the variational inequality.

We then obtain

Proposition 4 *Let $\hat{p}_F \in H^1(\widehat{\Omega}_s \cup \widehat{\mathscr{C}})$ be given. Find vector-valued velocities, vector-valued displacements, a scalar-valued fluid pressure, and a scalar-valued phase-field function, that is to say that $\{\hat{v}_f, \hat{v}_s, \hat{u}_f, \hat{u}_s, \hat{p}_f, \hat{\varphi}_s\} \in \{\hat{v}_f^D + \hat{V}^0_{f,\hat{v}}\} \times \hat{L}_s \times \{\hat{u}_f^D + \hat{V}^0_{f,\hat{u}}\} \times \{\hat{u}_s^D + \hat{V}^0_s\} \times \hat{L}^0_f \times H^1(\widehat{\Omega}_s \cup \widehat{\mathscr{C}})$, such that $\hat{v}_f(0) = \hat{v}_f^0$, $\hat{v}_s(0) = \hat{v}_s^0$, $\hat{u}_f(0) = \hat{u}_f^0, \hat{u}_s(0) = \hat{u}_s^0$ and $\hat{\varphi}_s(0) = \hat{\varphi}_s^0$ are satisfied, and for almost all times $t \in (0, T]$*

holds:

$$\textit{Fluid momentum}\begin{cases}(\hat{J}\hat{\rho}_f\partial_t\hat{v}_f,\hat{\psi}^v)_{\widehat{\Omega}_f}+(\hat{\rho}_f\hat{J}(\widehat{F}^{-1}(\hat{v}_f-\hat{w})\cdot\hat{\nabla})\hat{v}_f,\hat{\psi}^v)_{\widehat{\Omega}_f}\\ +(\hat{J}\hat{\sigma}_f\widehat{F}^{-T},\hat{\nabla}\hat{\psi}^v)_{\widehat{\Omega}_f}\\ +\langle\rho_f\nu_f\hat{J}(\widehat{F}^{-T}\hat{\nabla}\hat{v}_f^T\hat{n}_f)\widehat{F}^{-T},\hat{\psi}^v\rangle_{\widehat{\Gamma}_{out}}=0\quad\forall\hat{\psi}^v\in\hat{V}^0_{f,\widehat{\Gamma}_i},\end{cases}$$

$$\textit{Solid momentum, 1st eq.}\begin{cases}(\hat{\rho}_s\partial_t\hat{v}_s,\hat{\psi}^v)_{\widehat{\Omega}_s}+\Big(\big((1-\kappa)\tilde{\varphi}^2+\kappa\big)\widehat{F}\widehat{\Sigma},\hat{\nabla}\hat{\psi}^v\Big)_{\widehat{\Omega}_s}\\ +(\tilde{\varphi}^2\hat{p}_F,\hat{\nabla}\cdot\hat{\psi}^v)\quad=0\quad\forall\hat{\psi}^v\in\hat{V}^0_s,\end{cases}$$

$$\textit{Fluid mesh motion}\Big\{(\hat{\sigma}_{mesh},\hat{\nabla}\hat{\psi}^u)_{\widehat{\Omega}_f}\quad=0\quad\forall\hat{\psi}^u\in\hat{V}^0_{f,\hat{u},\widehat{\Gamma}_i},$$

$$\textit{Solid momentum, 2nd eq.}\Big\{\hat{\rho}_s(\partial_t\hat{u}_s-\hat{v}_s,\hat{\psi}^u)_{\widehat{\Omega}_s}\quad=0\quad\forall\hat{\psi}^u\in\hat{L}_s,$$

$$\textit{Fluid mass conservation}\Big\{(\widehat{div}\,(\hat{J}\widehat{F}^{-1}\hat{v}_f),\hat{\psi}^p)_{\widehat{\Omega}_f}\quad=0\quad\forall\hat{\psi}^p\in\hat{L}^0_f.$$

$$\textit{Phase-field}\begin{cases}(1-\kappa)(\hat{J}\hat{\varphi}_s\,\widehat{\Sigma}:\widehat{E},\hat{\psi}^\varphi)_{\widehat{\Omega}_s}+2(\hat{J}\hat{\varphi}_s\,\hat{p}_F\,\hat{\nabla}\cdot\hat{u}_s,\hat{\psi}^\varphi)_{\widehat{\Omega}_s}\\ +G_c\Big(-\frac{1}{\varepsilon}(\hat{J}(1-\hat{\varphi}_s),\hat{\psi}^\varphi)+\varepsilon(\hat{J}(\hat{\nabla}\hat{\varphi}_s\widehat{F}^{-1})\widehat{F}^{-T},\hat{\nabla}\hat{\psi}^\varphi)\Big)_{\widehat{\Omega}_s}\\ +\big(\hat{J}[\Xi+\gamma\partial_t\hat{\varphi}_s]^+,\hat{\psi}^\varphi\big)_{\widehat{\Omega}_s}\quad=0\quad\forall\hat{\psi}^\varphi\in H^1(\widehat{\Omega}_s\cup\widehat{\mathscr{C}}).\end{cases}$$

Remark 5 The continuous penalization constraint $[\Xi+\gamma\partial_t\hat{\varphi}_s]^+$ with $\Xi\in L^2$ and $\gamma>0$ and $[x]^+=\max(0,x)$ is numerically realized based on the incremental formulation as explained in [18].

Remark 6 The system in Proposition 4 has been implemented as presented. However the numerical example below deals with moderate deformations; namely $\|\hat{\nabla}\hat{u}\|\ll 1$ and thus $\widehat{F}\approx\hat{I}$ and $\hat{J}\approx 1$ such that the phase-field fracture model can be classified as proposed in [9, 15] and where linear elastic fracture mechanics applies. Testing large FSI-solid deformations, including fractures, is subject of ongoing studies.

Remark 7 (Modeling hypothesis) We emphasize that the given $\hat{p}_F$ in the pressurized phase-field fracture framework is neither coupled to the Navier-Stokes pressure $\hat{p}_f$ nor do we allow that the fracture reaches the FSI interface $\widehat{\Gamma}_i$. Mathematical modeling of these processes has not yet been established.

3 Aspects of Discretization and the Solution Algorithm

The coupled FSI phase-field problem in Proposition 4 is first formulated in terms of single semilinear form and then solved with the Rothe method: first time, then space. Specifically, time discretization is based on a One-step-θ scheme (here $\theta=0.5+\delta$, where $\delta = 0.01$ is related to the time step size $k = t^{n+1} - t^n$; resulting in A-stable (second order) stabilized Crank-Nicolson time-stepping) as presented for the pure FSI problem, Proposition 2, in [19]. Computational stability of these schemes has been investigated in [17]. In space, the problem is discretized with conforming finite elements on a quadrilateral mesh. For the fluid, an inf-sup stable velocity-pressure pair is chosen, Q_2^c/P_1^{dc}; for the displacements Q_2^c and for the phase-field variable Q_1^c. The fully-coupled nonlinear problem is solved with Newton's method as explained for pure FSI in detail in [19]. In particular, the Jacobian is constructed from the analytical evaluation of the directional derivatives. As linear solver we use UMFPACK [7], which is motivated by the fact that the numerical example is 2D and that preconditioners for the fully-coupled system are extremely difficult to be developed and out of scope in this contribution.

4 A Prototype Numerical Example

The example is computed with the finite element package deal.II [1] by implementing the method from [18] in the open source FSI code related to [19].

Configuration Details on the geometry can be found in Fig. 1. Here, the initial fracture is initialized by the initial condition $\hat{\varphi}^0 = 0$ in $(0.875, 0.9375) \times (0.25, 0.625)$. Three mesh levels are obtained from uniform refinement resulting in 2048, 8192 and 32,768 mesh cells. For the upper, lower, and left boundaries, the 'no-slip' conditions for velocity and no zero displacement for the solid are given. At the fluid outlet $\hat{\Gamma}_{out}$, the 'do-nothing' outflow condition [11] is imposed. A parabolic inflow velocity profile is given on $\hat{\Gamma}_{in}$ by

$$v_f(0, y) = \bar{U}(y-1)(y-0.5), \quad \bar{U} = 1.0\,\mathrm{m s}^{-1}.$$

For $t < 2.0\,\mathrm{s}$, $v_f(0, y)$ is scaled with $\frac{1-\cos(\frac{\pi}{2}t)}{2}$ in order to have a smooth inflow profile.

Parameters For the fluid we use $\varrho_f = 1\,\mathrm{kg\,m}^{-3}$, $\nu_f = 10^{-2}\,\mathrm{m}^2\,\mathrm{s}^{-1}$ resulting in stationary flow. The elastic solid is characterized by $\varrho_s = 10\,\mathrm{kg\,m}^{-3}$, $\nu_s = 0.2$, $\mu_s = 1\,\mathrm{kg\,m}^{-1}\,\mathrm{s}^{-2}$. The fracture pressure is $p_F = 10^{-2}\,\mathrm{Pa}$. The model parameter $\varepsilon = 0.044 = h_{coarse}$ is fixed in all computations as well as $\kappa = 10^{-10}$. Furthermore, $\gamma = 50$ and $G_c = 1\,\mathrm{N/m}$. The (absolute) Newton tolerance is chosen as 10^{-10}. Three augmented Lagrangian steps are performed per time step. The time step size is $k = 1\,\mathrm{s}$ and the total time $T = 10\,\mathrm{s}$.

Quantities of Interest We evaluate $\hat{u}_x(1, 0.75)$, the normal stress in x-direction (i.e., drag) along the FSI-interface, and a line integral, i.e., COD $= \int_{\{0\le x\le 2; y=0.4375\}} \hat{u}\hat{\nabla}\hat{\varphi}\, d\hat{s}$. Here, COD is related to the crack opening displacement in pure fracture problems. In addition, we check $\min(\hat{J}) > 0$.

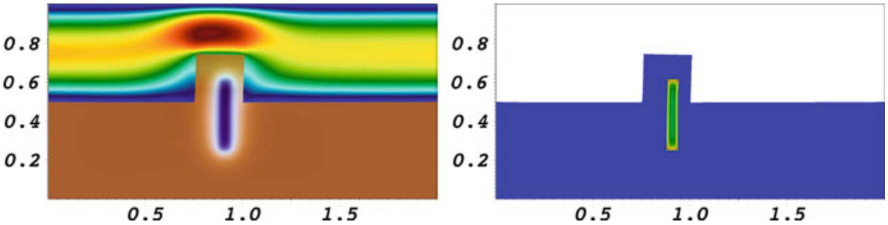


Fig. 2 At *left*: velocity field including the fracture pattern (in *blue*) in the *brown solid zone*. On the *right*, the penalty function Ξ is shown. The penalty is specifically active at the fracture boundary and the four corners (*red*). In the *blue* and *white* zones, we have $\Xi = 0$

Table 1 Goal functional evaluations on three different meshes at $T = 10$ s

Cells	DoFs	h[m]	$\hat{u}_x[\times 10^{-2}\text{m}]$	Drag$[\times 10^{-3}\text{N}]$	COD$[\times 10^{-4}\text{ m}]$	$\min(\hat{J})$
2048	41,829	0.044	1.851	2.995	03.339	0.787
8192	165,573	0.022	1.979	2.858	11.447	0.717
32, 768	658,821	0.011	2.077	2.771	15.293	0.621

Discussion of Our Findings Our results are provided in Fig. 2 and Table 1. Therein, we observe computational convergence of the first three goal functionals. Furthermore, $\min(\hat{J}) \gg 0$ showing that the ALE mapping is well-defined.

5 Conclusions

In this work, fluid-structure interaction has been coupled with a phase-field model for pressurized-fractures. The proposed framework is formulated in a variational-monolithic setting that ensures high accuracy of the coupling conditions. The emphasis in this work was on the model statement and a prototype numerical example. Therefore, the phase-field methodology has been rather used to represent a stationary pressurized fracture but not yet a propagating crack. Current work is based on adapting the different flow, solid, and fracture parameters to carry out fully nonstationary fluid-structure interaction with given and propagating fractures.

References

1. W. Bangerth, R. Hartmann, G. Kanschat, Deal.II – a general purpose object oriented finite element library. ACM Trans. Math. Softw. **33**(4), 24/1–24/27 (2007)
2. M.J. Borden, C.V. Verhoosel, M.A. Scott, T.J.R. Hughes, C.M. Landis, A phase-field description of dynamic brittle fracture. Comput. Methods Appl. Mech. Eng. **217**, 77–95 (2012)
3. B. Bourdin, Numerical implementation of the variational formulation for quasi-static brittle fracture. Interfaces Free Bound. **9**, 411–430 (2007)

4. B. Bourdin, G. Francfort, J.-J. Marigo, Numerical experiments in revisited brittle fracture. J. Mech. Phys. Solids **48**(4), 797–826 (2000)
5. B. Bourdin, G. Francfort, J.-J. Marigo, The variational approach to fracture. J. Elast. **91**(1–3), 1–148 (2008)
6. B. Bourdin, C. Larsen, C. Richardson, A time-discrete model for dynamic fracture based on crack regularization. Int. J. Frac. **168**(2), 133–143 (2011)
7. T.A. Davis, I.S. Duff, An unsymmetric-pattern multifrontal method for sparse LU factorization. SIAM J. Matrix Anal. Appl. **18**(1), 140–158 (1997)
8. J. Donea, S. Giuliani, J. Halleux, An arbitrary lagrangian-eulerian finite element method for transient dynamic fluid-structure interactions. Comput. Methods Appl. Mech. Eng. **33**, 689–723 (1982)
9. G. Francfort, J.-J. Marigo, Revisiting brittle fracture as an energy minimization problem. J. Mech. Phys. Solids **46**(8), 1319–1342 (1998)
10. T. Heister, M.F. Wheeler, T. Wick, A primal-dual active set method and predictor-corrector mesh adaptivity for computing fracture propagation using a phase-field approach. Comput. Methods Appl. Mech. Eng. **290**, 466–495 (2015)
11. J.G. Heywood, R. Rannacher, S. Turek, Artificial boundaries and flux and pressure conditions for the incompressible Navier-Stokes equations. Int. J. Numer. Methods Fluids **22**, 325–352 (1996)
12. J. Hron, S. Turek, A monolithic FEM/multigrid solver for an ALE formulation of fluid-structure interaction with applications in biomechanics, in *Fluid-Structure Interaction: Modelling, Simulation, Optimisation*, ed. by H.-J. Bungartz, M. Schäfer (Springer, Berlin/Heidelberg, 2006), pp. 146–170
13. T. Hughes, W. Liu, T. Zimmermann, Lagrangian-Eulerian finite element formulation for incompressible viscous flows. Comput. Methods Appl. Mech. Eng. **29**, 329–349 (1981)
14. C.J. Larsen, C. Ortner, E. Süli, Existence of solutions to a regularized model of dynamics fracture. Methods Appl. Sci. **20**, 1021–1048 (2010)
15. C. Miehe, F. Welschinger, M. Hofacker, Thermodynamically consistent phase-field models of fracture: variational principles and multi-field fe implementations. Int. J. Numer. Methods Eng. **83**, 1273–1311 (2010)
16. A. Mikelić, M.F. Wheeler, T. Wick, A quasi-static phase-field approach to pressurized fractures. Nonlinearity **28**(5), 1371–1399 (2015)
17. T. Richter, T. Wick, On time discretizations of fluid-structure interactions, in *Multiple Shooting and Time Domain Decomposition Methods*, ed. by T. Carraro, M. Geiger, S. Körkel, R. Rannacher. Contributions in Mathematical and Computational Science (Springer, 2015), pp. 377–400, http://www.springer.com/us/book/9783319233208
18. M. Wheeler, T. Wick, W. Wollner, An augmented-Lagangrian method for the phase-field approach for pressurized fractures. Comput. Methods Appl. Mech. Eng. **271**, 69–85 (2014)
19. T. Wick, Fluid-structure interactions using different mesh motion techniques. Comput. Struct. **89**(13–14), 1456–1467 (2011)

Weighted FEM for Two-Dimensional Elasticity Problem with Corner Singularity

Viktor A. Rukavishnikov

Abstract In this paper we consider homogeneous Dirichlet problem for the Lamé system with singularity caused by the reentrant corner to the boundary of the two-dimensional domain. For this problem we define the solution as a R_ν-generalized one; we state its existence and uniqueness in the weighted set $\overset{\circ}{\mathbf{W}}{}^{1}_{2,\nu}(\Omega,\delta)$. On the basis of the R_ν-generalized solution we construct weighted finite element method. We prove that the approximate solution converges to the exact one with the rate $O(h)$ in the norm of $\mathbf{W}^{1}_{2,\nu}(\Omega)$, and results of numerical experiments are presented.

1 Introduction

The generalized solution of the boundary value problem for the Lamé system in a two-dimensional domain with a boundary containing a reentrant corner γ belongs to the space $W_2^{1+\alpha-\varepsilon}(\Omega)$, where $0.25 \leq \alpha \leq 0.63$ for $\frac{3\pi}{2} \leq \gamma \leq 2\pi$ and ε is any positive number (see, e.g., [5]). Therefore, the approximate solution produced by classical finite element or finite difference schemes converges to a generalized solution no faster than at an $O(h^\alpha)$ rate.

By using special methods for extracting the singular part of the solution near corner points or applying grids refined toward the singularity point, it is possible to construct first-order accurate finite-element schemes (see, e.g., [1, 4, 5, 19, 21]).

Below, to construct a finite element method (FEM) without loss of accuracy for an elasticity problem in a domain with reentrant angles, the solution is determined as an R_ν-generalized one (see, e.g., [7–11, 15]). For boundary value problems with strongly singular solutions such that Dirichlet integral of the solution is divergent, it was shown in [2, 12–14, 16–18] that the FEM solution converges with first order of accuracy in space to the R_ν-generalized solution in the norms of weighted Sobolev and Lebesgue spaces. Below, for the R_ν-generalized solution of the Lamé system,

V.A. Rukavishnikov (✉)
Russian Academy of Sciences, Kim U Chen str. 65, Khabarovsk 680000, Russia

Far Eastern State Transport University, Khabarovsk, Russia
e-mail: vark0102@mail.ru

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_39

we proved an $O(h)$ convergence rate estimate independent of the reentrant angle on the boundary of the domain. The theoretical estimate is illustrated by numerical results obtained for a series of model problems.

2 Notation: Auxiliary Statements

Let R^2 denote the two-dimensional Euclidean space with elements $x = (x_1, x_2)$, $\|x\|^2 = x_1^2 + x_2^2$. Let $\Omega \subset R^2$ be a bounded non-convex polygonal domain with the boundary Γ containing one reentrant corner such that its vertex is located in the origin $O(0,0)$, $\bar{\Omega} = \Omega \cup \Gamma$.

We denote by $\Omega' = \{x \in \Omega : (x_1^2 + x_2^2)^{1/2} \leq \delta < 1\}$ a part of the δ-neighborhood of the point $(0,0)$ laying in the Ω. We introduce a weight function $\rho(x)$ that coincides in $\bar{\Omega}'$ with the distance to the origin, i.e. $\rho(x) = (x_1^2 + x_2^2)^{1/2}$ for $x \in \bar{\Omega}'$, and equals to δ for $x \in \bar{\Omega} \backslash \bar{\Omega}'$.

We introduce the weighted space $W_{2,\beta}^l(\Omega)$ with the squared norm

$$\|u\|_{W_{2,\beta}^l(\Omega)}^2 = \sum_{|i| \leq l} \|\rho^\beta |D^i u|\|_{L_2(\Omega)}^2, \tag{1}$$

where $D^i = \partial^{|i|}/\partial x_1^{i_1} \partial x_2^{i_2}$, $|i| = i_1 + i_2$, β is a nonnegative real number, l is a nonnegative integer. For $l = 0$ we have $W_{2,\beta}^0(\Omega) = L_{2,\beta}(\Omega)$.

By $W_{2,\alpha+l-1}^l(\Omega,\delta)$ for nonnegative real α, $l = 1, 2$, we denote the set of functions satisfying following conditions:

(a) $|D^m u(x)| \leq C_1 \left(\delta/\rho(x)\right)^{\alpha+m}$ for $x \in \bar{\Omega}'$, where $m = 0, \ldots, l$, $C_1 > 0$ is a constant independent of m,

(b) $\|u\|_{L_{2,\alpha}(\Omega \backslash \Omega')} \geq C_2 > 0$, $C_2 = \text{const}$; with the squared norm (1).

Let $L_{2,\alpha}(\Omega,\delta)$ be the set of functions satisfying conditions (a) and (b) with the squared norm (1) for $l = 0$.

The set $\overset{\circ}{W}{}_{2,\alpha+l-1}^l(\Omega,\delta) \subset W_{2,\alpha+l-1}^l(\Omega,\delta)$ $(l = 1, 2)$ is defined as the closure in norm (1) of the set $C_0(\Omega,\delta)$ of infinitely differentiable and finite in Ω functions satisfying conditions (a) and (b).

For the corresponding sets of vector-functions we use notation $\mathbf{W}_{2,\beta}^l(\Omega)$, $\mathbf{L}_{2,\beta}(\Omega)$, $\mathbf{W}_{2,\alpha+l-1}^l(\Omega,\delta)$, $\mathbf{L}_{2,\alpha}(\Omega,\delta)$, $\overset{\circ}{\mathbf{W}}{}_{2,\alpha+l-1}^l(\Omega,\delta)$, and $\mathbf{W}_2^l(\Omega)$ for non-weighted Sobolev vector-function spaces, $l = 1, 2$.

Lemma 1 ([7])

[A] Let $u \in W_{2,\alpha}^1(\Omega,\delta)$, then $\rho^\alpha u \in W_{2,0}^1(\Omega,\delta)$ and

$$|\rho^\alpha u|_{W_{2,0}^1(\Omega)} \leq C_3 \|u\|_{W_{2,\alpha}^1(\Omega)}, \textit{ where } C_3 = \text{const} > 0 \textit{ doesn't depend on } u;$$

[B] Let $\rho^\alpha u \in W^1_{2,0}(\Omega,\delta)$*, then* $u \in W^1_{2,\alpha}(\Omega,\delta)$ *and*

$$\|u\|_{W^1_{2,\alpha}(\Omega)} \le C_4\|\rho^\alpha u\|_{W^1_{2,0}(\Omega)},\ \textit{where}\ C_4 = \text{const} > 0\ \textit{doesn't depend on u;}$$

Lemma 2 ([10]) *If u belongs to* $W^2_{2,\alpha+1}(\Omega,\delta)$*, then there exists constant* C_5 *independent of u such that*

$$|\rho^{\alpha+1}u|^2_{W^2_{2,0}(\Omega)} \le C_5\|u\|^2_{W^2_{2,\alpha+1}(\Omega)}. \tag{2}$$

3 Problem Statement: R_ν-Generalized Solution

Let Ω be a homogeneous isotropic domain. In Ω we consider the boundary value problem for the displacement field $\mathbf{u} = (u_1, u_2)$ for the Lamé system with constant coefficients λ and μ:

$$-(2\,\mathbf{div}(\mu\varepsilon(\mathbf{u})) + \nabla(\lambda\,\mathrm{div}\,\mathbf{u})) = \mathbf{f}, \qquad x \in \Omega, \tag{3}$$

$$\mathbf{u} = \mathbf{0}, \qquad x \in \Gamma, \tag{4}$$

where $\varepsilon(\mathbf{u})$ is the strain tensor with components $\varepsilon_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right)$.

Assume that the right-hand side of (3) satisfies the condition

$$\mathbf{f} \in \mathbf{L}_{2,\beta}(\Omega,\delta), \quad \beta \ge 0. \tag{5}$$

Definition 3 Vector-function $\mathbf{u}_\nu = (u_{\nu,1}, u_{\nu,2})$ from the set $\mathring{\mathbf{W}}^1_{2,\nu}(\Omega,\delta)$ is called an R_ν-generalized solution to the problem (3), (4), if for every $\mathbf{v}$ from $\mathring{\mathbf{W}}^1_{2,\nu}(\Omega,\delta)$ the integral identity

$$a_\nu(\mathbf{u}_\nu, \mathbf{v}) = l_\nu(\mathbf{v}) \tag{6}$$

holds for any fixed value of $\nu \ge \beta$.

Here

$$a_\nu(\mathbf{u}_\nu,\mathbf{v}) = \int_\Omega \left(2\mu\varepsilon(\mathbf{u}_\nu) : \varepsilon(\rho^{2\nu}\mathbf{v}) + \lambda\,\mathrm{div}\,\mathbf{u}_\nu\,\mathrm{div}(\rho^{2\nu}\mathbf{v})\right)dx,\ \ l_\nu(\mathbf{v}) = \int_\Omega \rho^{2\nu}\mathbf{f}\cdot\mathbf{v}dx,$$

are bilinear and linear forms respectively.

Theorem 4 ([15]) *Let condition (5) be satisfied. Then for any $\nu > \beta$ there always exists parameter δ such that R_ν-generalized solution $\mathbf{u}_\nu$ to the problem (3), (4) exists and is unique in the set $\overset{\circ}{\mathbf{W}}{}^1_{2,\nu}(\Omega,\delta)$. In this case*

$$\|\mathbf{u}_\nu\|_{\mathbf{W}^1_{2,\nu}(\Omega)} \le C_6\|\mathbf{f}\|_{\mathbf{L}_{2,\beta}(\Omega)},$$

where C_6 is a positive constant independent of **f**.

Theorem 5 ([15]) *If for some δ there is a set of values ν such that R_ν-generalized solution to the problem (3), (4), and (5) exists in the set $\overset{\circ}{\mathbf{W}}{}^1_{2,\nu}(\Omega,\delta)$ then this solution is unique for all such ν.*

It is known that generalized solution to the problem (3), (4) can be represented as follows (see [20]):

$$\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \sum_{i=1}^{I} A_i r^{\alpha_i}\mathbf{s}_i^+(\theta) + \mathbf{u}^{reg}, \quad \mathbf{s}_i^+(\theta) = \begin{pmatrix} s_1^+(\theta) \\ s_2^+(\theta) \end{pmatrix}_i. \tag{7}$$

Here $\mathbf{s}_i^+(\theta)$ are the eigenfunctions of displacement, α_i are the corresponding eigenvalues, A_i are the stress intensity factors, $\mathbf{u}^{reg} \in \mathbf{W}_2^2(\Omega)$. For the considered problem minimal eigenvalue α_1 ranges in $[0.5, 0.62)$ for the domain with reentrant corner of magnitude from $3\pi/2$ to 2π (see [6]). It is obvious, that seminorm $|\mathbf{u}|_{W^2_{2,\nu}(\Omega)}$ is finite for all $\nu > 1-\alpha_1$ and $\mathbf{u} \in \mathbf{W}^2_{2,\nu}(\Omega,\delta)$.

Let us consider integral equality (6). By Lemma 1 functions $\rho^{2\nu}\mathbf{v}$ belong to the set $\mathbf{W}^1_{2,0}(\Omega,\delta)$. We substitute generalized solution of the problem (3), (4) into (6). Derived equation is true for the definition of a weak solution. Hence, generalized solution is an R_ν-generalized solution too.

By Theorem 4 R_ν-generalized solution exists and is unique for $\nu > \beta$, by Theorem 5 it is the same for all such ν. Therefore R_ν-generalized solution coincides with the generalized one and $\mathbf{u}_\nu \in \mathbf{W}^2_{2,\nu}(\Omega,\delta)$ for $\nu > \max\{1-\alpha_1, \beta\}$.

4 Weighted Finite Element Method

On basis of the R_ν-generalized solution for the posed problem we construct a scheme of the weighted finite element method (see [14]). With this purpose, we perform a quasi-uniform triangulation of the domain Ω such that conventional requirements are satisfied (see e.g. [17]). An approximate R_ν-generalized solution to the problem (3), (4), and (5) has the form $\mathbf{u}_\nu^h = (u^h_{\nu,1}, u^h_{\nu,2})$,

$$u^h_{\nu,1} = \sum_{i=0}^{n-1} d_{2i}\psi^i(x),\; u^h_{\nu,2} = \sum_{i=0}^{n-1} d_{2i+1}\psi^i(x),\; d_j = \rho^{-\nu^*}(P_{[j/2]})c_j,\; j = \overline{0, 2n-1}.$$

Here $\psi^i(x) = \rho^{\nu^*}(x)\varphi^i(x)$ is a basis function associated with inner node P_i ($i = \overline{0, n-1}$); $\varphi^i(x)$ is linear on each finite element and $\varphi^i(P_j) = \delta_{ij}$, $i, j = \overline{0, n-1}$, δ_{ij} is a Kronecker delta, ν^* is a real number, h is the maximum size of sides of triangular elements K of the domain decomposition. The set V^h is defined as the linear span of the system of basis functions $\{\psi^i\}_{i=0}^{n-1}$, $\mathbf{V}^h = [V^h]^2$.

Definition 6 Vector-function $\mathbf{u}_\nu^h \in \mathbf{V}^h$ is called an approximate R_ν-generalized solution to the problem (3), (4), and (5) produced by the weighted finite element method if for every $\mathbf{v}^h \in \mathbf{V}^h$ the integral identity

$$a(\mathbf{u}_\nu^h, \mathbf{v}^h) = l(\mathbf{v}^h)$$

holds for any fixed value of $\nu \geq \beta$.

We construct an interpolant $\mathbf{u}_\nu^I = (u_{\nu,1}^I, u_{\nu,2}^I)$

$$u_{\nu,k}^I = \sum_{i=0}^{n-1} \rho^\nu(P_i) u_{\nu,k}(P_i)\psi^i(x), \quad k = 1, 2, \quad \nu > 0, \quad \nu^* = -\nu.$$

Lemma 7 *Let $\mathbf{u}_\nu$ be an R_ν-generalized solution to the problem* (3), (4), *and* (5) *and $\mathbf{u}_\nu^h$ be an approximate R_ν-generalized solution by the weighted FEM. Then there exist positive constants C_7, C_8 independent of the set $\mathbf{V}^h$ such that*

$$\|\mathbf{u}_\nu - \mathbf{u}_\nu^h\|_{\mathbf{W}_{2,\nu}^1(\Omega)} \leq C_7 \inf_{\mathbf{v}^h \in \mathbf{V}^h} \|\mathbf{u}_\nu - \mathbf{v}^h\|_{\mathbf{W}_{2,\nu}^1(\Omega)} \leq C_8 \|\mathbf{u}_\nu - \mathbf{u}_\nu^I\|_{\mathbf{W}_{2,\nu}^1(\Omega)}.$$

The proof of this lemma follows from the Céa's lemma.

Theorem 8 *Suppose that the R_ν-generalized solution to the problem* (3), (4), *and* (5) *belongs to $\mathbf{W}_{2,\nu}^2(\Omega, \delta)$. Then there exists a constant C_9 independent of $\mathbf{u}_\nu$ and h such that the following estimation holds*

$$\|\mathbf{u}_\nu - \mathbf{u}_\nu^h\|_{\mathbf{W}_{2,\nu}^1(\Omega)} \leq C_9 h \|\mathbf{u}_\nu\|_{\mathbf{W}_{2,\nu}^2(\Omega)} \tag{8}$$

for the constructed triangulation of the domain Ω.

Proof Let us estimate $\|u_{\nu,k} - u_{\nu,k}^I\|_{W_{2,\nu}^1(\Omega)}$ for $k = 1, 2$. For brevity, we introduce the notation $u_{\nu,k} = u$, $u_{\nu,k}^I = u_I$.

Functions u, u_I belong to $W_{2,\nu}^1(\Omega, \delta)$. Then by Lemma 1 $\rho^\nu u$, $\rho^\nu u_I$ belong to $W_{2,0}^1(\Omega, \delta)$ and

$$\|u - u_I\|_{W_{2,\nu}^1(\Omega)} \leq C_4 \|\rho^\nu u - \rho^\nu u_I\|_{W_{2,0}^1(\Omega)}. \tag{9}$$

Consider the following norm for an arbitrary element K of the triangulation

$$\|\rho^{\nu}u-\rho^{\nu}u_I\|_{W^1_{2,0}(K)} \tag{10}$$

We notice that for single element $\rho^{\nu}u_I=\sum\limits_{i=1}^{3}\rho^{\nu}(P_i^K)u(P_i^K)\varphi_i(x)$, where P_i^K are the vertices of the triangle K, $i=1,2,3$ are those local numbers, $\varphi_i(P_j^K)=\delta_{ij}$, $j=1,2,3$. By theorem 2 from the [3] the following estimation holds

$$\|\rho^{\nu}u-\rho^{\nu}u_I\|_{W^1_{2,0}(K)}\le C_{10}h|\rho^{\nu}u|_{W^2_{2,0}(K)}, \tag{11}$$

where constant C_{10} does not depend on h and u.

Summing (11) over all elements, using Lemma 2 and estimation (9), we obtain

$$\begin{aligned}\|u-u_I\|_{W^1_{2,\nu}(\Omega)}&\le C_4\|\rho^{\nu}u-\rho^{\nu}u_I\|_{W^1_{2,0}(\Omega)}\le\\&\le C_4C_{10}h|\rho^{\nu}u|_{W^2_{2,0}(\Omega)}\le C_3C_4C_{10}h\|u\|_{W^2_{2,\nu}(\Omega)}.\end{aligned} \tag{12}$$

Taking into account that

$$\|\mathbf{u}_{\nu}-\mathbf{u}_{\nu}^I\|^2_{\mathbf{W}^1_{2,\nu}(\Omega)}=\|u_{\nu,1}-u^I_{\nu,1}\|^2_{W^1_{2,\nu}(\Omega)}+\|u_{\nu,2}-u^I_{\nu,2}\|^2_{W^1_{2,\nu}(\Omega)},$$

and using estimation (12) we obtain

$$\|\mathbf{u}_{\nu}-\mathbf{u}_{\nu}^I\|_{\mathbf{W}^1_{2,\nu}(\Omega)}\le C_{11}h\|\mathbf{u}_{\nu}\|_{\mathbf{W}^2_{2,\nu}(\Omega)},$$

where constant C_{11} is defined as maximum values of $C_3C_4C_{10}$ for each component.

By the latter inequality and Lemma 7 we get (8) with constant $C_9=C_8C_{11}$.

5 Results of Numerical Experiments

In this section we present results of numerical solution of the non-homogeneous Dirichlet problem for the Lamé system in the domain $\Omega=(-1,1)\times(-1,1)\setminus[0,1]\times[-1,0]\subset R^2$.

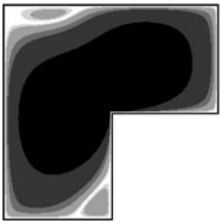
As a solution to the model problem we choose vector-function $\mathbf{u}$ with components containing both singular and regular components, regular part belongs to $\mathbf{W}^2_2(\Omega)$:

$$\begin{aligned}u_1&=\cos(x_1)\cos^2(x_2)(x_1^2+x_2^2)^{0.3051}+(x_1^2+x_2^2),\\u_2&=\cos^2(x_1)\cos(x_2)(x_1^2+x_2^2)^{0.3051}+(x_1^2+x_2^2).\end{aligned}$$

Table 1 Dependence of relative errors of the generalized (η) and R_ν-generalized (η_ν) ($\delta = 0.0029$, $\nu = 1.2$, $\nu^* = 0.16$) solution of the problem on the mesh step

$2N$	128		256		512		1024		2048		4096
h	1.105e−2		5.524e−3		2.762e−3		1.381e−3		6.905e−4		3.453e−4
η	2.849e−2	1.54	1.850e−2	1.53	1.205e−2	1.53	7.870e−3	1.53	5.146e−3	1.53	3.367e−3
η_ν	2.868e−2	1.57	1.827e−2	1.65	1.107e−2	2.16	5.117e−3	2.21	2.319e−3	1.98	1.171e−3

Table 2 Number, percentage equivalence, and distribution of nodes where absolute error $|e_1|$ (generalized solution) is not less than given limit values

	$\lvert e_1\rvert$	Limit	$\lvert e_1\rvert$	
$2N$	Distributionl	values	%	Number
4096		● ≥ 5e−6	48.078	6045622
		● ≥ 1e−6	29.387	3695278
		● ≥ 5e−7	6.724	845466
		● ≥ 1e−7	9.624	1210158
		● ≥ 5e−8	2.564	322439
		○ ≥ 0	3.624	455758

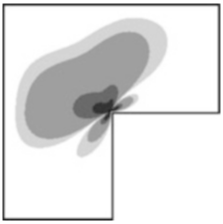
Calculations were performed by the program “Proba-IV” for different values of N (N is a half of number of partitioning segments along the greater side) with regular triangular meshes ($h = \sqrt{2}/N$). Generalized solution was determined by the integral equality (6) for $\nu = 0$.

In Table 1 we present values of relative errors of the generalized solution in the norm of the space $\mathbf{W}_2^1$ $\left(\eta = \frac{\|\mathbf{e}\|_{\mathbf{W}_2^1}}{\|\mathbf{u}\|_{W_2^1}}\right)$, and the R_ν-generalized one in the norm of the weighted space $\mathbf{W}_{2,\nu}^1$ $\left(\eta_\nu = \frac{\|\mathbf{e}_\nu\|_{\mathbf{W}_{2,\nu}^1}}{\|\mathbf{u}\|_{W_{2,\nu}^1}}\right)$ with different values of h. Here $\mathbf{e} = (e_1, e_2) = (u_1 - u_1^h, u_2 - u_2^h)$ and $\mathbf{e}_\nu = (e_{\nu,1}, e_{\nu,2}) = (u_1 - u_{\nu,1}^h, u_2 - u_{\nu,2}^h)$. In addition Table 1 contains ratios between error norms, obtained on meshes with step reducing twice. Tables 2 and 3 contain for giving limit values: number of nodes where $|e_1|$, $|e_{\nu,1}|$ belong to the giving range; this number in percentage to the total number of nodes; pictures of the absolute error distribution in the domain Ω.

Thus, we have found the following:

1. an approximate R_ν-generalized solution of the problem converges to the exact one with the rate $O(h)$ in the norm of the space $\mathbf{W}_{2,\nu}^1(\Omega)$ in contrast with the generalized one which converges with the rate $O(h^{0.61})$ for the classical FEM;
2. for the approximate R_ν-generalized solution obtained by the weighted finite element method, an absolute error value is by one or two orders of magnitude less than for the approximate generalized one obtained by the FEM.

Table 3 Number, percentage equivalence, and distribution of nodes where absolute error $|e_{\nu,1}|$ (R_ν-generalized solution, $\delta = 0.0029$, $\nu = 1.2$, $\nu^* = 0.16$) is not less than given limit values

	$\lvert e_1 \rvert$	Limit	$\lvert e_1 \rvert$	
$2N$	Distribution|	values	%	Number
4096		● $\geq$ 5e−6 ● $\geq$ 1e−6 ● $\geq$ 5e−7 ● $\geq$ 1e−7 ● $\geq$ 5e−8 ○ $\geq$ 0	0.033 0.771 2.481 21.789 12.588 62.339	4108 96899 311996 2739862 1582876 7838980

Acknowledgements This work was made under the frame of project 16-11-10008 of the RSCF.

References

1. T. Apel, A.-M. Sändig, J.R. Whiteman, Graded mesh refinement and error estimates for finite element solutions of elliptic boundary value problems in non-smooth domains. Math. Methods Appl. Sci. **19**(1), 63–85 (1996)
2. A.Y. Bespalov, V.A. Rukavishnikov, The exponential rate of convergence of the finite-element method for the Dirichlet problem with singularity of the solution. Dokl. Math. **62**(2), 266–270 (2000)
3. J.H. Bramble, M. Zlámal, Triangular elements in the finite element method. Math. Comput. **24**(112), 809–820 (1970)
4. P. Morin, R.H. Hochetto, K.G. Siebert, Convergence of adaptive finite element methods. SIAM Rev. **44**(5), 631–658 (2002)
5. H. Nguyen-Xuan, R.G. Liu, S. Bordas, S. Natarajan, T. Rabczuk, An adaptive singular ES-FEM for mechanics problems with singular field of arbitrary order. Comput. Methods Appl. Mech. Eng. **253**, 252–273 (2013)
6. A. Rössle, Corner singularities and regulatiy of weak solutions for the two-dimensional Lamé equations on domains with angular corners. J. Elast. **60**(1), 57–75 (2000)
7. V.A. Rukavishnikov, On the differential properties of R_ν-generalized solution of Dirichlet problem. Dokl. Akad. Nauk SSSR **309**(6), 1318–1320 (1989)
8. V.A. Rukavishnikov, The Dirichlet problem for a second-order elliptic equation with noncoordinated degeneration of the input data. Differ. Equ. **32**(3), 406–412 (1996)
9. V.A. Rukavishnikov, On the existence and uniqueness of an R_ν-generalized solution of a boundary value problem with uncoordinated degeneration of the input data. Dokl. Math. **90**(2), 562–564 (2014)
10. V.A. Rukavishnikov, E.V. Kuznetsova, Coercive estimate for a boundary value problem with noncoordinated degeneration of the data. Differ. Equ. **43**(4), 550–560 (2007)
11. V.A. Rukavishnikov, E.V. Kuznetsova, The R_ν-generalized solution of a boundary value problem with a singularity belongs to the space $W^{k+2}_{2,\nu+\beta/2+k+1}(\Omega, \delta)$. Differ. Equ. **45**(6), 913–917 (2009)
12. V.A. Rukavishnikov, A.O. Mosolapov, New numerical method for solving time-harmonic Maxwell equations with strong singularity. J. Comput. Phys. **231**(6), 2438–2448 (2012)

13. V.A. Rukavishnikov, A.O. Mosolapov, Weighted edge finite element method for Maxwell's equations with strong singularity. Dokl. Math. **87**(2), 156–159 (2013)
14. V.A. Rukavishnikov, S.G. Nikolaev, Weighted finite element method for an elasticity problem with singularity. Dokl. Math. **88**(3), 705–709 (2013)
15. V.A. Rukavishnikov, S.G. Nikolaev, On the R_{ν}-generalized solution of the Lamé system with corner singularity. Dokl. Math. **92**(1), 421–423 (2015)
16. V.A. Rukavishnikov, H.I. Rukavishnikova, Finite-element method for the 1st boundary-value problem with the coordinated degeneration of the initial data. Dokl. Akad. Nauk **338**(6), 731–733 (1994)
17. V.A. Rukavishnikov, H.I. Rukavishnikova, The finite element method for a boundary value problem with strong singularity. J. Comput. Appl. Math. **234**(9), 2870–2882 (2010)
18. V.A. Rukavishnikov, H.I. Rukavishnikova, On the error estimation of the finite element method for the boundary value problems with singularity in the Lebesgue weighted space. Numer. Funct. Anal. Optim. **34**(12), 1328–1347 (2013)
19. B. Szabó, I. Babuška, *Finite Element Analysis* (Wiley, New York, 1991)
20. Z. Yosibash, *Singularities in Elliptic Boundary Value Problems and Elasticity and Their Connection with Failure Initiation*. Intersisciplinary Applied Mathematics, vol. 37 (Springer, New York, 2012)
21. O.C. Zienkiewicz, R.L. Taylor, J.Z. Zhu, *The Finite Element Method: Its Basis and Fundamentals*, 6th edn. (Elsevier, Oxford/Boston, 2005)

A Local Error Estimate for the Poisson Equation with a Line Source Term

Tobias Köppl, Ettore Vidotto, and Barbara Wohlmuth

Abstract In this paper, we show a local a priori error estimate for the Poisson equation in three space dimensions (3D), where the source term is a Dirac measure concentrated on a line. This type of problem can be found in many application areas. In medical engineering, e.g., blood flow in capillaries and tissue can be modeled by coupling Poiseuille's and Darcy's law using a line source term. Due to the singularity induced by the line source term, finite element solutions converge suboptimal in classical norms. However, quite often the error at the singularity is either dominated by model errors (e.g. in dimension reduced settings) or is not the quantity of interest (e.g. in optimal control problems). Therefore we are interested in local error estimates, i.e., we consider in space a L^2-norm on a fixed subdomain excluding a neighborhood of the line, where the Dirac measure is concentrated. It is shown that linear finite elements converge optimal up to a log-factor in such a norm. The theoretical considerations are confirmed by some numerical tests.

1 Introduction

Our model problem is defined on an open, convex and polyhedral domain $\Omega \subset \mathbb{R}^3$. Within Ω we consider a $\mathscr{C}^2$-curve Γ, having the following properties:

$$\Gamma \subset \Omega, \qquad |\Gamma| \leq C < \infty \qquad \text{and} \qquad \text{dist}(\Gamma, \partial\Omega) \geq c > 0, \tag{1}$$

This work was partially supported by the German Research Foundation (DFG) under the contract/grant numbers: WO-671/11-1.

T. Köppl
Institut für Wasser- und Umweltsystemmodellierung, Universität Stuttgart, D-70569 Stuttgart, Germany
e-mail: tobias.koeppl@iws.uni-stuttgart.de

E. Vidotto (✉) • B. Wohlmuth
Institute for Numerical Mathematics, Technische Universität München, D-85748 Garching b. München, Germany
e-mail: vidotto@ma.tum.de; wohlmuth@ma.tum.de

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_40

where $C, c > 0$ are fixed constants and dist denotes the Euclidean distance. Using these definitions, the model problem reads as:

$$\begin{cases} -\Delta u = \delta_\Gamma & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{2}$$

where δ_Γ is the Dirac measure concentrated on the curve Γ. Considering a more general Poisson problem with a source term that is given by a real and regular Borel measure, one can show existence and uniqueness of a weak solution [3, Thm. 1]. Moreover it can be proved that it belongs to the space $W_0^{1,p}(\Omega)$ for $p \in [1, 3/2)$ and that linear finite element solutions have a reduced convergence order of $1/2$ with respect to the standard L^2-norm.

For the case of a Dirac measure concentrated on a $\mathscr{C}^2$-curve fulfilling the conditions (1), the authors of [9] proved that the weak solution of (2) is in $W_0^{1,p}(\Omega)$ for $p \in [1, 2)$ (see [9, Thm. 2.1, Case (ii)]) and that linear finite elements converge with first order in the standard L^2-norm.

In [5], the authors have performed some numerical tests to investigate the convergence behavior of an elliptic 3D-1D coupled problem arising in the context of blood flow simulations [4, 7]. Thereby, a 1D problem is defined on a straight line in 3D and is embedded into the 3D problem by a Dirac source term concentrated on this line. The L^2-error of the 3D problem has not been computed on Ω, but on a subdomain $\Omega \setminus \tilde{Z}_R$, where $\tilde{Z}_R$ is a cylinder of radius R around this line. It has been observed that the local L^2-convergence behavior of linear finite elements is optimal up to a log-factor.

Motivated by these numerical results, we prove a quasi-optimal convergence behavior with respect to a local L^2-norm. For this purpose, we define a domain Z_R covering a certain neighborhood of the curve Γ:

$$Z_R := \{x \in \Omega : \mathrm{dist}(x, \Gamma) < R\},$$

where $0 < R < \mathrm{dist}(\Gamma, \partial\Omega)$ is a fixed constant. The rest of this paper is organized as follows: In Sect. 2, we formulate the main result, which is a quasi-optimal bound of the L^2-error on $\Omega \setminus Z_R$ and give an outline of the proof. A key ingredient for this proof is an auxiliary result, which is proved in Sect. 3. Finally in Sect. 4, we illustrate our theoretical result by some numerical tests.

2 Main Result

The weak formulation of problem (2) reads as follows: Find $u \in W_0^{1,p}(\Omega)$ for $p \in [1, 2)$ such that

$$(\nabla u, \nabla \varphi) = \int_\Gamma \varphi \, d\Gamma \qquad \forall \varphi \in W_0^{1,q}(\Omega), \tag{3}$$

where $(\cdot,\cdot)$ denotes the duality pairing of $L^p(\Omega)$ and $L^q(\Omega)$, and $q > 2$ satisfies $1/p + 1/q = 1$. The right-hand side of the weak problem (3) is well-defined, since for $q > 3$ the embedding $W^{1,q}(\Omega) \hookrightarrow \mathscr{C}(\Omega)$ is continuous (see, e.g., [1, Thm. 4.12]).

In order to solve (3), we introduce a family of quasi-uniform simplicial triangulations $\mathscr{T}_h = \{\tau\}$ of Ω, where h_τ is the diameter of the element τ. Furthermore, let $h := \max_{\tau\in\mathscr{T}_h} h_\tau$ be the meshsize. Our finite element space is then defined by standard conforming linear finite elements:

$$V_h := \{\varphi_h \in H_0^1(\Omega) : \varphi_{h|\tau} \in \mathbb{P}_1(\tau),\ \forall \tau \in \mathscr{T}_h\}, \tag{4}$$

where $\mathbb{P}_1(\tau)$ denotes the space of linear polynomials on the element τ. Due to $V_h \subset W_0^{1,q}(\Omega) \subset W_0^{1,p}(\Omega)$, the following discrete version of (3) is well-defined: Find $u_h \in V_h$ such that

$$(\nabla u_h, \nabla \varphi_h) = \int_\Gamma \varphi_h \, d\Gamma \qquad \forall \varphi_h \in V_h. \tag{5}$$

In order to derive an upper bound for the finite element error, we use a standard duality argument and define the corresponding dual problem as follows:

$$\begin{cases} -\Delta w = e\chi_{\Omega\setminus Z_R} & \text{in } \Omega, \\ \quad w = 0 & \text{on } \partial\Omega, \end{cases} \tag{6}$$

where $\chi_{\Omega\setminus Z_R}$ is the characteristic function of $\Omega \setminus Z_R$ and $e := u - u_h$. Since Ω is convex and $e\chi_{\Omega\setminus Z_R} \in L^2(\Omega)$, it holds that $w \in H^2(\Omega) \cap H_0^1(\Omega)$ (see [8, Chap. 8]).

From now on, we fix $p = 7/5$. The choice is somehow arbitrary, but guarantees that $H^2(\Omega) \subset W^{1,q}(\Omega)$ for $q = 7/2$. Moreover, we have $w \in W^{1,\frac{7}{2}}(\Omega)$. A weak formulation of (6) is given by:

$$(\nabla w, \nabla \varphi) = (e\chi_{\Omega\setminus Z_R}, \varphi), \quad \forall \varphi \in W^{1,\frac{7}{5}}(\Omega) \tag{7}$$

and it is well defined. The corresponding finite element approximation $w_h \in V_h$ satisfies the following equality:

$$(\nabla w_h, \nabla \varphi_h) = (e\chi_{\Omega\setminus Z_R}, \varphi_h), \quad \forall \varphi_h \in V_h. \tag{8}$$

For the proof of our main result, a pointwise error estimate for the finite element error $w - w_h$ is required. Such an estimate is presented in the following lemma.

Lemma 1 *Let $w \in H_0^1(\Omega) \cap H^2(\Omega)$ be the solution of problem (6) and let $w_h \in V_h$ be its finite element approximation given by (8). Then for a $x_0 \in \Gamma$, the following pointwise error estimate holds*

$$|(w - w_h)(x_0)| \lesssim h^2 |\ln h| \|w\|_{H^2(\Omega)}. \tag{9}$$

We use the notation $\lesssim$ for $\leq C$, with a generic constant C independent of h. The proof of this lemma is presented in the next section. Comparing this result with the standard L^∞-estimate [6, Thm. 22.7], we have on the right-hand side of (9) the norm $\|w\|_{H^2(\Omega)}$ and not the stronger norm $\|w\|_{W^{2,\infty}(\Omega)}$. This is a consequence of the H^2-regularity of w and the fact that w is harmonic in Z_R. By means of this lemma, one can prove the following theorem, which is the main result of our paper.

Theorem 2 *Let $u \in W^{1,p}(\Omega)$ for $p \in [1,2)$ be the weak solution of (2) and let $u_h \in V_h$ be its finite element approximation. Then it holds the following quasi-optimal estimate for the local error:*

$$\|u-u_h\|_{L^2(\Omega\backslash Z_R)} \lesssim h^2|\ln h|. \tag{10}$$

Proof Using the Galerkin orthogonality, properties (1) of Γ, Lemma 1 and H^2-regularity of w, we obtain:

$$\begin{aligned}\|e\|^2_{L^2(\Omega\backslash Z_R)} &= (e\chi_{\Omega\backslash Z_R}, u-u_h) = (\nabla w, \nabla(u-u_h))\\ &= (\nabla(w-w_h), \nabla(u-u_h)) = (\nabla u, \nabla(w-w_h))\\ &= \int_\Gamma (w-w_h)d\Gamma \leq \int_\Gamma |w-w_h|d\Gamma \leq |\Gamma| \max_{x\in\Gamma} |(w-w_h)(x)|\\ &\lesssim h^2|\ln h|\ \|w\|_{H^2(\Omega)} \lesssim h^2|\ln h|\|e\|_{L^2(\Omega\backslash Z_R)}.\end{aligned}$$

3 Proof of Lemma 1

Let $x_0 \in \Gamma$ be fixed. We denote a ball of radius r around x_0 by B_r and assume $h < r/4$. Moreover, we demand $B_{r+2h} \subset B_{2r} \subset Z_R$. In order to study the finite element error of the dual problem within B_r, a smooth cut-off function η is introduced. Besides $\eta \in \mathscr{C}^\infty(\Omega)$, this function is supposed to satisfy:

$$\begin{cases} \eta(x) = 1, & \text{if } x \in B_r,\\ \eta(x) = 0, & \text{if } x \in \Omega \setminus B_{2r},\\ 0 \leq \eta(x) \leq 1, & \text{if } x \in B_{2r} \setminus B_r.\end{cases}$$

By the help of this cut-off function, an auxiliary Dirac problem is introduced, where the Dirac measure is concentrated at $x_0 \in \Gamma$:

$$\begin{cases} -\Delta z = \alpha(\delta_{x_0} + f_0) & \text{in } \Omega,\\ z = 0 & \text{on } \partial\Omega,\end{cases} \tag{11}$$

with $\alpha := \operatorname{sgn}\big((w - w_h)(x_0)\big)$. The function f_0 is defined as follows:

$$f_0(x) := \begin{cases} -2\nabla\eta \cdot \nabla G_{x_0} - G_{x_0}\Delta\eta & \text{if } x \in B_{2r} \setminus B_r \\ 0 & \text{if } x \notin B_{2r} \setminus B_r \end{cases}$$

and $G_{x_0}(x) := \frac{1}{4\pi}\frac{1}{|x-x_0|}$ denotes Green's function in 3D with respect to x_0 for $x \neq x_0$. This choice of f_0 leads to the following solution of the Dirac problem (11):

$$z = \alpha \cdot \eta \cdot G_{x_0}.$$

A straightforward computation shows that $z \notin H^1(\Omega)$, but $z \in W_0^{1,p}(\Omega)$ for $p \in [1, 3/2)$. The weak formulation of (11) reads now: Find $z \in W_0^{1,\frac{7}{5}}(\Omega)$ such that:

$$(\nabla z, \nabla\varphi) = \alpha\left(\int_\Omega f_0\varphi d\Omega \;+\; \varphi(x_0)\right),\ \forall\varphi \in W_0^{1,\frac{7}{2}}(\Omega). \tag{12}$$

The discrete version of (12) is given by:

$$(\nabla z_h, \nabla\varphi_h) = \alpha\left(\int_\Omega f_0\varphi_h\, d\Omega \;+\; \varphi_h(x_0)\right),\ \forall\varphi_h \in V_h. \tag{13}$$

Setting $\varphi = w - w_h$, we obtain by (12):

$$(\nabla z, \nabla(w - w_h)) = (f_0, \alpha(w - w_h)) + |(w - w_h)(x_0)|.$$

This yields:

$$|(w - w_h)(x_0)| = (\nabla z, \nabla(w - w_h)) - (f_0, \alpha(w - w_h)). \tag{14}$$

Using the Hölder inequality, $f_0 \in L^2(\Omega)$, $w \in H^2(\Omega)$ and standard finite element estimates, one obtains obviously:

$$\int_\Omega \alpha(w - w_h)f_0\, d\Omega \leq \|f_0\|_{L^2(\Omega)} \cdot \|w - w_h\|_{L^2(\Omega)} \lesssim h^2.$$

Inserting this bound in (14) and using the Galerkin orthogonality, it follows:

$$|(w - w_h)(x_0)| \lesssim h^2 + (\nabla(z - z_h), \nabla(w - w_h)).$$

It remains to estimate the second term on the right-hand side. For this purpose, we consider an interpolation operator $S_h : W^{n,p}(\Omega) \to V_h$ of Scott-Zhang type [11], due to its stability properties. By Galerkin orthogonality, we have:

$$|(w-w_h)(x_0)| \lesssim h^2 + (\nabla(z-z_h), \nabla(w-S_h w)).$$

Splitting the domain Ω into B_r and $\Omega \setminus B_r$ and applying the Hölder inequality yields:

$$\begin{aligned} |(w-w_h)(x_0)| \lesssim h^2 &+ \|\nabla(z-z_h)\|_{L^2(\Omega\setminus B_r)} \|\nabla(w-S_h w)\|_{L^2(\Omega\setminus B_r)} \\ &+ \ \|\nabla(z-z_h)\|_{L^1(B_r)} \|\nabla(w-S_h w)\|_{L^\infty(B_r)}. \end{aligned} \tag{15}$$

Next, we have to bound the four error terms occuring on the right-hand side of (15). The estimates for these error terms are provided in the following.

For the second term $\|\nabla(w-S_h w)\|_{L^2(\Omega\setminus B_r)}$, a standard estimate (see, e.g. [11, Thm 4.1]) and the H^2-regularity of w, yields:

$$\|\nabla(w-S_h w)\|_{L^2(\Omega\setminus B_r)} \lesssim h\|w\|_{H^2(\Omega)}, \tag{16}$$

Next we estimate the interpolation error $\|\nabla(w-S_h w)\|_{L^\infty(B_r)}$ using some results about the interior regularity of w.

Lemma 3 *Let w be the solution of (7) and let us assume that $B_{r+2h} \subset B_{2r} \subset Z_R$ and $4h < r$ holds. Then we have on B_r:*

$$\|\nabla(w-S_h w)\|_{L^\infty(B_r)} \lesssim h\|w\|_{H^2(\Omega)}. \tag{17}$$

Proof Using the approximation properties of S_h and the Sobolev embedding $W^{4,2}(B_{r+2h}) \hookrightarrow W^{2,\infty}(B_{r+2h})$ [1, Thm. 4.12], one obtains:

$$\|\nabla(w-S_h w)\|_{L^\infty(B_r)} \lesssim h\|w\|_{W^{2,\infty}(B_{r+2h})} \lesssim h\,\|w\|_{W^{4,2}(B_{r+2h})}.$$

Interior regularity [8, Thm. 8.10], H^2-regularity of w and the fact that w is harmonic in Z_R (and in particular in B_{2r}), yield:

$$\|w\|_{W^{4,2}(B_{r+2h})} \lesssim \|w\|_{W^{1,2}(B_{2r})} \lesssim \|w\|_{H^2(\Omega)},$$

which completes the proof.

Finally, we have to derive suitable bounds for the error terms involving the finite element error $z-z_h$. Since z is the weak solution of a homogeneous Poisson problem whose source term consists of a L^2-function and a pointwise Dirac measure, we can use the results derived in [10].

Lemma 4 *Let $z \in H^2(\Omega \setminus B_r) \cap W_0^{1,p}(\Omega)$ for $p \in [1, 3/2)$ be defined by (12) and let $z_h \in V_h$ be its finite element approximation defined by (13). Then the following two bounds are valid:*

$$\|\nabla(z-z_h)\|_{L^2(\Omega\setminus B_r)} \lesssim h + \|z-z_h\|_{L^2(\Omega\setminus B_{r-2h})}. \tag{18}$$

and

$$\|\nabla(z - z_h)\|_{L^1(B_r)} \lesssim h|\ln h|. \tag{19}$$

Proof The proof of this lemma is a direct consequence of [10, Lemma 3.6] and [10, Lemma 3.10].

Combining equation (16) and Lemmas 3 and 4, we find

$$|(w - w_h)(x_0)| \lesssim h^2|\ln h|\|w\|_{H^2(\Omega)} + h\|z - z_h\|_{L^2(\Omega\setminus B_{r-2h})}\|w\|_{H^2(\Omega)}.$$

Using [10, Thm. 2.1] for the term $\|z - z_h\|_{L^2(\Omega\setminus B_{r-2h})}$ and the H^2-regularity of w, it finally follows

$$|(w - w_h)(x_0)| \lesssim h^2|\ln h|\|w\|_{H^2(\Omega)}. \tag{20}$$

4 Numerical Experiments

In this section, we illustrate the theoretical estimate (10) by numerical examples. For this purpose, we choose the unit cube as our computational domain, i.e., $\Omega = (0, 1)^3$. By the help of the curves:

$$\Gamma_1 : [0, 1] \to \overline{\Omega}, \quad \Gamma_1(s) = (0.5, 0.5, s),$$
$$\Gamma_2 : [0, 1] \to \Omega, \quad \Gamma_2(s) = 0.5\,(1 + 0.06\cos(2\pi s)\,, 1 + 0.06\sin(2\pi s), s + 0.5)\,,$$

we define two different Poisson problems, setting in (2) Γ to Γ_1 or Γ_2 .

In the first case, we modify the homogeneous boundary condition in (2), such that the exact solution of the considered Poisson problem is given by:

$$u(x) = -\frac{1}{2\pi}\log|x - \Gamma_1(x_3)|.$$

In the second case, we keep the homogeneous boundary condition. However, for this problem no analytical solution is available. Therefore, we precompute a numerical reference solution on a fine mesh and determine the finite element errors by comparing numerical solutions on coarser meshes to the reference solution. The numerical results (see Fig. 1) are obtained by means of a linear finite element solver implemented in DUNE [2].

According to Theorem 2, we report the local L^2-error on $\Omega \setminus Z_R$ for different radii $R \in \{0, 0.0625, 0.125, 0.25\}$. The approximation errors and convergence rates for the different refinement levels ℓ are listed in Tables 1 and 2. It can be seen that for $R = 0$ the finite element scheme converges only with first order, as predicted by

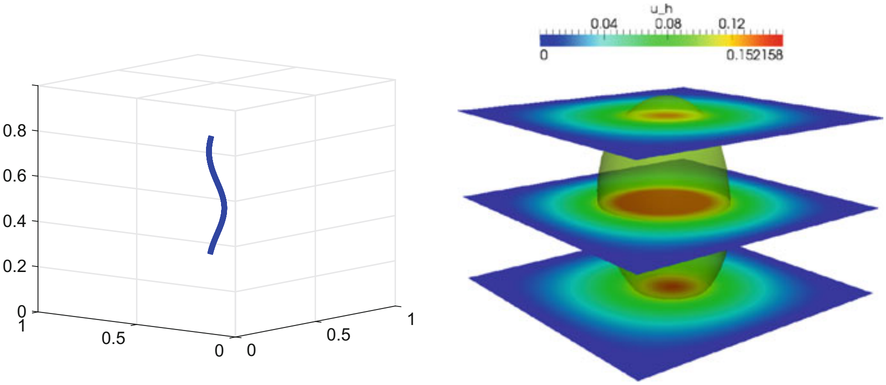


Fig. 1 The figure shows the position of Γ_2 within the unit cube (*left*) and a contour plot for $u_h = 0.1$ together with three different slices at $x_3 = 0.2, 0.5, 0.8$ (*right*)

Table 1 L^2-error on $\Omega \setminus Z_R$ for $\Gamma = \Gamma_1$

ℓ	Dof	$R = 0$	Rate	$R = 0.0625$	Rate	$R = 0.125$	Rate	$R = 0.25$	Rate
2	125	1.0461e−2	/	1.0461e−2	/	1.0461e−2	/	4.3424e−3	/
3	729	5.4648e−3	**0.94**	5.4648e−3	0.94	2.5806e−3	2.02	1.4692e−3	1.56
4	4913	2.7741e−3	**0.98**	1.3369e−3	2.03	8.0692e−4	1.68	2.3815e−4	**2.63**
5	35,937	1.3968e−3	**0.99**	4.1809e−4	1.68	1.4183e−4	**2.51**	5.7466e−5	**2.05**
6	274,625	7.0077e−4	**1.00**	7.4890e−5	**2.48**	3.4535e−5	**2.04**	1.4533e−5	**1.98**
7	2,146,689	3.5097e−4	**1.00**	1.8260e−5	**2.04**	8.7158e−6	**1.99**	3.6356e−6	**2.00**

Table 2 Discrete L^2-error on $\Omega \setminus Z_R$ for $\Gamma = \Gamma_2$

ℓ	Dof	$R = 0$	Rate	$R = 0.0625$	Rate	$R = 0.125$	Rate	$R = 0.25$	Rate
2	125	9.9816e−3	/	5.7708e−3	/	4.1503e−3	/	2.8019e−3	/
3	729	6.5546e−3	0.61	2.6437e−3	1.13	1.5200e−3	1.45	6.3984e−4	**2.13**
4	4913	4.6708e−3	0.49	1.1080e−3	1.25	3.9200e−4	**1.96**	1.5896e−4	**2.01**
5	35,937	1.8626e−3	**1.33**	2.8450e−4	**1.96**	1.0354e−4	**1.92**	3.9731e−5	**2.00**
6	274,625	7.7970e−4	**1.26**	5.6445e−5	**2.33**	2.1513e−5	**2.27**	8.2663e−6	**2.26**

[9, Lemma 3.3]. In the remaining cases second order convergence can be observed, if $h < R$ holds. This is in agreement with estimate (10) in Theorem 2.

References

1. R. Adams, J. Fournier, *Sobolev Spaces*, vol. 140 (Academic press, Amsterdam/Boston, 2003)
2. P. Bastian et al., A generic grid interface for parallel and adaptive scientific computing. Part I: abstract framework. Computing **82**(2–3), 103–119 (2008)
3. E. Casas, *L^2 estimates for the finite element method for the Dirichlet problem with singular data*. Numer. Math. **47**(4), 627–632 (1985)

4. L. Cattaneo, P. Zunino, A computational model of drug delivery through microcirculation to compare different tumor treatments. Int. J. Numer. Methods Biomed. Eng. **30**, 1347–1371 (2014). Wiley Online Library
5. L. Cattaneo, P. Zunino, *Numerical Investigation of Convergence Rates for the FEM Approximation of 3D-1D Coupled Problems*. Numerical Mathematics and Advanced Applications-ENUMATH 2013 (Springer International Publishing, Cham/Heidelberg/New York, 2015), pp. 727–734
6. P. Ciarlet, Basic error estimates for elliptic problems. Handb. Numer. Anal. **2**, 17–351 (1991)
7. C. D'Angelo, P. Zunino, *Multiscale Models of Drug Delivery by Thin Implantable Devices*. Applied and Industrial Mathematics in Italy III. Series on Advances in Mathematics for Applied Sciences (World Scientific, Singapore, 2009), pp. 298–310
8. D. Gilbarg, N. Trudinger, *Elliptic partial differential equations of second order* (Springer, Berlin/New York, 2015)
9. W. Gong, G. Wang, N. Yan, Approximations of elliptic optimal control problems with controls acting on a lower dimensional manifold. SIAM J. Control Optim. **52**(3), 2008–2035 (2014)
10. T. Köppl, B. Wohlmuth, Optimal a priori error estimates for an elliptic problem with Dirac right-hand side. SIAM J. Numer. Anal. **52**(4), 1753–1769 (2014)
11. R. Scott, S. Zhang, Finite element interpolation of nonsmooth functions satisfying boundary conditions. Math. Comput. **54**(190), 483–493 (1990)

Multirate Undrained Splitting for Coupled Flow and Geomechanics in Porous Media

Kundan Kumar, Tameem Almani, Gurpreet Singh, and Mary F. Wheeler

Abstract We consider a multirate iterative scheme for the quasi-static Biot equations modelling the coupled flow and geomechanics in a porous medium. The iterative scheme is based on undrained splitting where the flow and mechanics equations are decoupled with the mechanics solve followed by the pressure solve. The multirate scheme proposed here uses different time steps for the two equations, that is, uses q flow steps for each coarse mechanics step and may be interpreted as using a regularization parameter for the mechanics equation. We prove the convergence of the scheme and the proof reveals the appropriate regularization parameter and also the effect of the number of flow steps within coarse mechanics step on the convergence rate.

1 Introduction

Coupling of geomechanics and flow in poroelastic media has many important applications such as subsidence events, carbon sequestration, ground water remediation, hydrocarbon production, enhanced geothermal systems, solid waste disposal, and biomedical modeling. Starting from the pioneering work of Terzaghi and Biot [1], there has been active investigation into the coupled geomechanics and flow problems [11]. The Biot model consists of a geomechanics equation coupled to a flow model with the displacement, pressure and flow velocity as unknowns. There is a huge literature on Biot equations and they have been analyzed by a number of

K. Kumar (✉)
Mathematics Institute, University of Bergen, Bergen, Norway
e-mail: kundan.kumar@uib.no

T. Almani • G. Singh • M.F. Wheeler
CSM, ICES, UT Austin, USA
e-mail: tameem@ices.utexas.edu; gurpreet@ices.utexas.edu; mfw@ices.utexas.edu

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_41

authors who established existence, uniqueness, and regularity, see Showalter [13], Phillips and Wheeler [8] and references therein.

In contrast to a fully implicit scheme for solving the coupled model of Biot equations, iterative methods are often employed in practice [2–4]. The iterative schemes allow the decoupling of the flow and mechanics equations and thus offer several attractive features (such as use of existing flow and mechanics codes, use of appropriate pre-conditioners and solvers for the two models, and ease of implementation). The design of iterative schemes however is an important consideration for an efficient, convergent, and robust algorithm. In addition, often we can take a coarser time step for the mechanics equation than for the flow.

Here, we consider one of the iterative schemes often used in practice: undrained splitting and propose a multirate iterative scheme. This scheme considers a finer time step for the flow model and a coarser time step for mechanics (q flow steps for each mechanics step) and then performs an iteration between the mechanics and finer flow steps.The scheme is iterative in the sense that for each coarse mechanics time step, we solve for q flow finer time steps followed by a mechanics step and we further repeat the process. Details about convergence criteria can be found in [7]. The converged solutions solve the coupled time-discrete system consisting of q flow solves and one mechanics solve. The flow finer solve uses the mechanics at the coarse step and hence, the coupled system is fully implicit. Since the cost of mechanics is often much more than the flow, a less mechanics solves leads to computational savings. Our work is motivated by the recent work of Mikelić and Wheeler [6, 7] where they have considered different iterative schemes for flow and mechanics couplings and established contractive results in suitable norms, (see also [5] for studying the von Neumann stability of iterative algorithms, [12] for multirate schemes for Darcy-Stokes, and [9, 10] for relationship of these iterative methods to the linearization procedures).

2 Model Equations, Discretization and Splitting Algorithm

We assume a linear, elastic, homogeneous, and isotropic poro-elastic medium $\Omega \subset \mathbb{R}^d$, $d = 2$ or 3, in which the reservoir is saturated with a slightly compressible viscous fluid. The fluid is assumed to be slightly compressible and its density is a linear function of pressure, with a constant viscosity $\mu_f > 0$. The reference density of the fluid $\rho_f > 0$, the Lamé coefficients $\lambda > 0$ and $G > 0$, the dimensionless Biot coefficient α, and the pore volume φ^* are all positive. The absolute permeability tensor, $\boldsymbol{K}$, is assumed to be symmetric, bounded, uniformly positive definite in space and constant in time. A quasi-static Biot model [1] will be employed in this work.

The model reads: Find $\boldsymbol{u}$ and p satisfying the equations below for all time $t \in]0, T[$:

Flow Equation:

$$\frac{\partial}{\partial t}\Big(\big(\frac{1}{M} + c_f\varphi_0\big)p + \alpha\nabla\cdot\boldsymbol{u}\Big) - \nabla\cdot\Big(\frac{1}{\mu_f}\boldsymbol{K}\big(\nabla p - \rho_{f,r} g\nabla\eta\big)\Big) = \tilde{q} \text{ in } \Omega$$

Mechanics Equations: $-\operatorname{div}\boldsymbol{\sigma}^{\text{por}}(\boldsymbol{u}, p) = \boldsymbol{f} \text{ in } \Omega,$

$$\boldsymbol{\sigma}^{\text{por}}(\boldsymbol{u}, p) = \boldsymbol{\sigma}(\boldsymbol{u}) - \alpha p \boldsymbol{I} \text{ in } \Omega,$$

$$\boldsymbol{\sigma}(\boldsymbol{u}) = \lambda(\nabla\cdot\boldsymbol{u})\boldsymbol{I} + 2G\boldsymbol{\varepsilon}(\boldsymbol{u}) \text{ in } \Omega$$

Boundary Cond.: $\boldsymbol{u} = \boldsymbol{0}\ ,\ \boldsymbol{K}(\nabla p - \rho_{f,r} g\nabla\eta)\cdot\boldsymbol{n} = 0 \text{ on } \partial\Omega$

Initial Cond. (t=0): $\Big(\big(\frac{1}{M} + c_f\varphi_0\big)p + \alpha\nabla\cdot\boldsymbol{u}\Big)(0) = \big(\frac{1}{M} + c_f\varphi_0\big)p_0 + \alpha\nabla\cdot\boldsymbol{u}_0.$

where: g is the gravitational constant, η is the distance in the direction of gravity (assumed to be constant in time), $\rho_{f,r} > 0$ is a constant reference density (relative to the reference pressure p_r), φ_0 is the initial porosity, M is the Biot constant, $\tilde{q} = \frac{q}{\rho_{f,r}}$ where q is a mass source or sink term taking into account injection into or out of the reservoir. We remark that the above system is linear and coupled through the Biot coefficient terms (involving α).

2.1 Mixed Variational Formulation

We use a mixed finite element formulation for flow and a conformal Galerkin formulation for mechanics for the spatial discretization and a backward-Euler for the time discretization. Let $\mathfrak{T}_h$ denote a regular family of conforming triangular elements of the domain of interest, $\overline{\Omega}$. Using the lowest order Raviart-Thomas (RT) spaces , we have the following discrete spaces ($\boldsymbol{V}_h$ for discrete displacements, Q_h for discrete pressures, and $\boldsymbol{Z}_h$ for discrete velocities (fluxes)):

$$\boldsymbol{V}_h = \{\boldsymbol{v}_h \in H^1(\Omega)^d\ ;\ \forall T \in \mathfrak{T}_h, \boldsymbol{v}_{h|T} \in \mathbb{P}_1{}^d, \boldsymbol{v}_{h|\partial\Omega} = \boldsymbol{0}\},$$

$$Q_h = \{p_h \in L^2(\Omega)\ ;\ \forall T \in \mathfrak{T}_h, p_{h|T} \in \mathbb{P}_0\},$$

$$\boldsymbol{Z}_h = \{\boldsymbol{q}_h \in H(\text{div}; \Omega)^d\ ; \forall T \in \mathfrak{T}_h, \boldsymbol{q}_{h|T} \in \mathbb{P}_1{}^d,\ \boldsymbol{q}_h\cdot\boldsymbol{n} = 0 \text{ on } \partial\Omega\}.$$

We also assume that the finer time step is given by: $\Delta t = t_k - t_{k-1}$. If we denote the total number of timesteps by N, then the total simulation time is given by T = Δt N, and $t_i = i\Delta t$, $0 \leqslant i \leqslant N$ denote the discrete time points. The proof presented here can be easily extended to other mixed method approaches (e.g., [14]) or Conformal Galerkin discretizations. **Notation:** k denotes the coarser time step iteration index (for indexing mechanics coarse time steps), m is the finer (local) time step iteration index (for indexing flow fine time steps), Δt stands for the unit (finer) time step, and q is the “fixed” number of local flow time steps per coarse mechanics time step.

3 Multirate Formulation: Undrained Split Iterative Method

The scheme starts by solving the mechanics problem followed by a sequence of flow problems, and iterates between the two until convergence is achieved. The iteration assumes a constant fluid mass during the deformation of the structure (and can be interpreted as a regularization of mechanics equation). We start by presenting the scheme.

3.1 Undrained Split Multirate Algorithm

1. For k = 0, q, 2q, 3q .. (**Mechanics time step iteration level**)
2. – For n = 1,2, . . . (**coupling iteration index**)
   – **First Step: Mechanics equations**
   Given $p_h^{n,k+q} \in Q_h$ solve for $\boldsymbol{u}_h^{n+1,k+q} \in V_h$ satisfying:

$$-2G\nabla \cdot (\boldsymbol{\varepsilon}(\boldsymbol{u}_h^{n+1,k+q})) - (\lambda + L)\nabla \cdot \big((\nabla \cdot \boldsymbol{u}_h^{n+1,k+q})I\big) = -\alpha\nabla \cdot \big(p_h^{n,k+q} I\big) - L\nabla \cdot \big((\nabla \cdot \boldsymbol{u}_h^{n,k+q})I\big) + \boldsymbol{f} \tag{1}$$

   – **Second Step: Flow equations**

   (a) Given $\boldsymbol{u}_h^{n+1,k+q}$, for m = 1,2,..,q (**flow time step iteration level**)

   - solve for $p_h^{n+1,m+k} \in Q_h$ and $z_h^{n+1,m+k} \in \boldsymbol{Z}_h$ satisfying:

$$\beta\Big(\frac{p_h^{n+1,m+k} - p_h^{n+1,m-1+k}}{\Delta t}\Big) + \frac{1}{\mu_f}\nabla \cdot \boldsymbol{z}_h^{n+1,m+k} = -\alpha\nabla \cdot \Big(\frac{\boldsymbol{u}_h^{n+1,k+q} - \boldsymbol{u}_h^{k}}{q\Delta t}\Big) + \tilde{q}_h \tag{2}$$

$$\boldsymbol{z}_h^{n+1,m+k} = -\boldsymbol{K}\big(\nabla p_h^{n+1,m+k} - \rho_{f,r} g \nabla \eta\big) \tag{3}$$

In the above, we have used $\beta = \big(\frac{1}{M} + c_f\varphi_0\big)$ for the notational convenience. L is a regularization parameter and the corresponding term vanishes in the case of convergence.

4 Analysis of the Multirate Undrained Scheme

The convergence proof is based on studying the difference of two successive iterates and deriving the contraction of appropriate quantities in suitable norms. Accordingly, we define:

$$\delta\xi^{n,k} = \xi^{n+1,k} - \xi^{n,k}, \text{ where } \xi = p, \boldsymbol{u}, \text{or } \boldsymbol{z}.$$

It is interesting that the contracting quantity is a composite one consisting of both pressure $p^{n+1,k+m}$ and volumetric strain terms $\nabla \cdot \boldsymbol{u}^{n+1,k+q}$. For a particular coupling iteration, $n \geq 1$, and between two coarse mechanics time steps t_k and t_{k+q}, we define the quantity to be contracted on as:

$$m^{n+1,k+m} = \frac{L}{\gamma q}\nabla \cdot \boldsymbol{u}^{n+1,k+q} + \frac{\alpha}{\gamma}(p^{n+1,k+m} - p^{n+1,k+m-1}), \text{ for } 1 \leq m \leq q,$$

where γ is an adjustable coefficient that will be selected carefully such that the scheme achieves contraction on m. The presence of γ does not alter the contractivity, however, it simplifies the algebra and provides a systematic technique for obtaining similar results for other problems.

4.1 Weak Formulation of Difference of Two Successive Iterates

Considering the difference between one local flow iteration and its corresponding local flow iteration in the previous coupling iteration, and the difference between two consecutive mechanics coupling iterations, the weak formulation of equations corresponding to (1), (2), and (3) can be written as follows.

Mechanics Step: Given $\delta p_h^{n,k+q}$ from the previous coupling iteration, find $\delta \boldsymbol{u}_h^{n+1,k+q} \in \boldsymbol{V}_h$ such that,

$$\forall \boldsymbol{v}_h \in V_h\,,\; 2G\big(\boldsymbol{\varepsilon}(\delta\boldsymbol{u}_h^{n+1,k+q}), \boldsymbol{\varepsilon}(\boldsymbol{v}_h)\big) + (\lambda + L)\big(\nabla \cdot \delta\boldsymbol{u}_h^{n+1,k+q}, \nabla \cdot \boldsymbol{v}_h\big) = \alpha\big(\delta p_h^{n,k+q}, \nabla \cdot \boldsymbol{v}_h\big) + L\big(\nabla \cdot \delta\boldsymbol{u}_h^{n,k+q}, \nabla \cdot \boldsymbol{v}_h\big)\,, \tag{4}$$

Flow Step: Given $\delta\boldsymbol{u}_h^{n+1,k+q}$, for $1 \leq m \leq q$, find $\delta p_h^{n+1,m+k} \in Q_h$, $\delta \boldsymbol{z}_h^{n+1,m+k} \in \boldsymbol{Z}_h$ such that:

$$\forall \theta_h \in Q_h\,,\; \beta\Big(\frac{\delta p_h^{n+1,m+k} - \delta p_h^{n+1,m-1+k}}{\Delta t}, \theta_h\Big) + \frac{1}{\mu_f}\big(\nabla \cdot \delta \boldsymbol{z}_h^{n+1,m+k}, \theta_h\big) = -\frac{\alpha}{q\Delta t}\Big(\nabla \cdot \big(\delta\boldsymbol{u}_h^{n+1,k+q} - \delta\boldsymbol{u}_h^{n+1,k}\big), \theta_h\Big) \tag{5}$$

$$\forall \boldsymbol{q}_h \in \boldsymbol{Z}_h\,,\; \big(\boldsymbol{K}^{-1}\delta \boldsymbol{z}_h^{n+1,m+k}, \boldsymbol{q}_h\big) = (\delta p_h^{n+1,m+k}, \nabla \cdot \boldsymbol{q}_h) \tag{6}$$

We note that $\delta \boldsymbol{u}_h^{n+1,k}$ in (5) is essentially zero, as the value of $\boldsymbol{u}_h^k$ is already computed from the previous multirate iterative coupling iteration. Therefore, it can be omitted.

Step 1: Mechanics Equation

First, we analyze the mechanics equation. Testing (4) with $\boldsymbol{v}_h = \delta \boldsymbol{u}_h^{n+1,k+q}$, we get:

$$\begin{aligned}
2G\left\|\boldsymbol{\varepsilon}(\delta \boldsymbol{u}_h^{n+1,k+q})\right\|^2 &+ (\lambda + L)\left\|\nabla \cdot \delta \boldsymbol{u}_h^{n+1,k+q}\right\|^2 \\
&= \left(\alpha \delta p_h^{n,k+q} + L\nabla \cdot \delta \boldsymbol{u}_h^{n,k+q}, \nabla \cdot \delta \boldsymbol{u}_h^{n+1,k+q}\right) \\
&= \left(\sum_{m=1}^{q}\left(\alpha\left(\delta p_h^{n,m+k} - \delta p_h^{n,m-1+k}\right) + \frac{L}{q}\nabla \cdot \delta \boldsymbol{u}_h^{n,k+q}\right), \nabla \cdot \delta \boldsymbol{u}_h^{n+1,k+q}\right) \\
&\leq \frac{\varepsilon}{2}\left\|\nabla \cdot \delta \boldsymbol{u}^{n+1,k+q}\right\|^2 + \frac{1}{2\varepsilon}\gamma^2 \sum_{m=1}^{q}\left\|\delta m^{n,k+m}\right\|^2
\end{aligned}$$

by noting that $\sum_{m=1}^{q}\left(\delta p_h^{n,m+k} - \delta p_h^{n,m-1+k}\right) = \delta p_h^{n,k+q}$ and using Young's inequality. For $\varepsilon = \lambda + L$, we obtain after some simplifications,

$$\frac{4G}{\lambda + L}\left\|\boldsymbol{\varepsilon}(\delta \boldsymbol{u}_h^{n+1,k+q})\right\|^2 + \left\|\nabla \cdot \delta \boldsymbol{u}_h^{n+1,k+q}\right\|^2 \leq \frac{\gamma^2}{(\lambda + L)^2}\sum_{m=1}^{q}\left\|\delta m^{n,k+m}\right\|^2. \tag{7}$$

Step 2: Flow Equation

Testing (5), with $\theta_h = \delta p_h^{n+1,m+k} - \delta p_h^{n+1,m-1+k}$, and multiplying by Δt, we get: (recall $\beta = \frac{1}{M} + c_f\varphi_0$)

$$\begin{aligned}
&\beta\left\|\delta p_h^{n+1,m+k} - \delta p_h^{n+1,m-1+k}\right\|^2 \\
&+ \frac{\Delta t}{\mu_f}\left(\nabla \cdot \delta z_h^{n+1,m+k}, \delta p_h^{n+1,m+k} - \delta p_h^{n+1,m-1+k}\right) = \\
&-\frac{\alpha}{q}\left(\nabla \cdot \delta \boldsymbol{u}_h^{n+1,k+q}, \delta p_h^{n+1,m+k} - \delta p_h^{n+1,m-1+k}\right)
\end{aligned} \tag{8}$$

Now, consider (6) for two consecutive local flow finer time steps, $t = t_{m+k}$, and $t = t_{m-1+k}$, and test with $\boldsymbol{q}_h = \delta z_h^{n+1,m+k}$ and taking the difference between them, we get

$$\left(\boldsymbol{K}^{-1}\left(\delta z_h^{n+1,m+k}-\delta z_h^{n+1,m-1+k}\right),\delta z_h^{n+1,m+k}\right)$$
$$=\left(\delta p_h^{n+1,m+k}-\delta p_h^{n+1,m-1+k},\nabla\cdot\delta z_h^{n+1,m+k}\right). \tag{9}$$

Substituting (9) into (8), we have

$$\beta\left\|\delta p_h^{n+1,m+k}-\delta p_h^{n+1,m-1+k}\right\|^2+\frac{\Delta t}{\mu_f}\left(\boldsymbol{K}^{-1}\left(\delta z_h^{n+1,m+k}-\delta z_h^{n+1,m-1+k}\right),\delta z_h^{n+1,m+k}\right)=$$
$$-\frac{\alpha}{q}\left(\nabla\cdot\delta\boldsymbol{u}_h^{n+1,k+q},\delta p_h^{n+1,m+k}-\delta p_h^{n+1,m-1+k}\right).$$

By Young's inequality, with further simplifications,

$$\beta\left\|\delta p_h^{n+1,m+k}-\delta p_h^{n+1,m-1+k}\right\|^2+\frac{\alpha}{q}\left(\nabla\cdot\delta\boldsymbol{u}_h^{n+1,k+q},\delta p_h^{n+1,m+k}-\delta p_h^{n+1,m-1+k}\right)$$
$$\frac{\Delta t}{2\mu_f}\left(\left\|\boldsymbol{K}^{-1/2}\delta z_h^{n+1,m+k}\right\|^2-\left\|\boldsymbol{K}^{-1/2}\delta z_h^{n+1,m-1+k}\right\|^2\right.$$
$$\left.+\left\|\boldsymbol{K}^{-1/2}(\delta z_h^{n+1,m+k}-\delta z_h^{n+1,m-1+k})\right\|^2\right)=0.$$

Summing for q local flow time steps and after some simplifications (telescopic cancellations together with the fact that $\delta z_h^{n+1,k}=0$), we get

$$\beta\sum_{m=1}^{q}\left(\left\|\delta p_h^{n+1,m+k}-\delta p_h^{n+1,m-1+k}\right\|^2+\frac{\alpha}{q}\left(\nabla\cdot\delta\boldsymbol{u}_h^{n+1,k+q},\delta p_h^{n+1,m+k}-\delta p_h^{n+1,m-1+k}\right)\right)$$
$$\frac{\Delta t}{2\mu_f}\left\|\boldsymbol{K}^{-1/2}\delta z_h^{n+1,k+q}\right\|^2+\frac{\Delta t}{2\mu_f}\sum_{m=1}^{q}\left\|\boldsymbol{K}^{-1/2}\left(\delta z_h^{n+1,m+k}-\delta z_h^{n+1,m-1+k}\right)\right\|^2=0. \tag{10}$$

Step 3: Combining Mechanics and Flow

Multiplying (10) by another free parameter c^2 and adding (10), we obtain

$$\frac{4G}{\lambda+L}\left\|\boldsymbol{\varepsilon}(\delta\boldsymbol{u}_h^{n+1,k+q})\right\|^2+\sum_{m=1}^{q}\left\{c^2\beta\left\|\delta p_h^{n+1,m+k}-\delta p_h^{n+1,m-1+k}\right\|^2\right.$$
$$\left.+\frac{c^2\alpha}{q}\left(\nabla\cdot\delta\boldsymbol{u}_h^{n+1,k+q},\delta p_h^{n+1,m+k}-\delta p_h^{n+1,m-1+k}\right)+\left\|\nabla\cdot\delta\boldsymbol{u}_h^{n+1,k+q}\right\|^2\right\}$$

$$+\frac{c^2\Delta t}{2\mu_f}\left\|\boldsymbol{K}^{-1/2}\delta z_h^{n+1,k+q}\right\|^2+\frac{c^2\Delta t}{2\mu_f}\sum_{m=1}^{q}\left\|\boldsymbol{K}^{-1/2}\left(\delta z_h^{n+1,m+k}-\delta z_h^{n+1,m-1+k}\right)\right\|^2$$

$$\leq\frac{\gamma^2}{(\lambda+L)^2}\sum_{m=1}^{q}\left\|\delta m^{n,k+m}\right\|^2. \qquad (11)$$

Step 4: Identifying the Parameters

Note that we have three free parameters: $c^2, \gamma,$ and L. Below we provide the procedure for determining these parameters yielding a contraction. These parameters should be chosen such that the terms on the left hand side of (11) remain positive, and the scheme achieves contraction on m. Clearly,

$$\left\|\delta m^{n+1,k+m}\right\|^2=\frac{L^2}{q^2\gamma^2}\left\|\nabla\cdot\delta\boldsymbol{u}^{n+1,k+q}\right\|^2+\frac{\alpha^2}{\gamma^2}\left\|(p^{n+1,k+m}-p^{n+1,k+m-1})\right\|^2$$
$$+\frac{2\alpha L}{\gamma^2 q}\left((p^{n+1,k+m}-p^{n+1,k+m-1}),\nabla\cdot\delta\boldsymbol{u}^{n+1,k+q}\right).$$

Matching coefficients by comparing with the terms in the curly brackets in (11) provides us $\frac{L^2}{q^2\gamma^2}=1$, $\frac{\alpha^2}{\gamma^2}\leq c^2\beta$, and $\frac{2\alpha L}{\gamma^2 q}=\frac{c^2\alpha}{q}$. This gives, $L=q\gamma,\quad L\geq\frac{\alpha^2}{2\beta}$ and since the contraction factor is monotone with respect to L, its minimum is achieved when we choose,

$$L=\frac{\alpha^2}{2\beta}\text{ implying }\gamma=\frac{\alpha^2}{2q\beta}\text{ and }c^2=\frac{4q^2\beta}{\alpha^2}.$$

Using above in (11) we note that the contraction factor is $\frac{L^2}{q^2(\lambda+L)^2}$ and is smaller when q is larger. Also, when $q=1$, the above contraction rate reduces to that of the single rate case [6] (when the time steps for the mechanics and flow are the same).

5 Main Result: Contraction

Our main result summarises the above contraction result.

Theorem 1 *With* $L = \dfrac{\alpha^2}{2\beta}$ *and* $c^2 = \frac{4q^2\beta}{\alpha^2}$, *the undrained multirate iterative scheme defined by* (1), (2), *and* (3) *is a contraction given by*

$$\frac{c^2 \Delta t}{2\mu_f} \left\| \boldsymbol{K}^{-1/2} \delta z_h^{n+1,k+q} \right\|^2 + \frac{c^2 \Delta t}{2\mu_f} \sum_{m=1}^{q} \left\| \boldsymbol{K}^{-1/2} \left(\delta z_h^{n+1,m+k} - \delta z_h^{n+1,m-1+k} \right) \right\|^2$$
$$+ \sum_{m=1}^{q} \left\| \delta m^{n+1,k+m} \right\|^2 + \frac{4G}{\lambda + L} \left\| \boldsymbol{\varepsilon}(\delta \boldsymbol{u}_h^{n+1,k+q}) \right\|^2 \leq \frac{L^2}{q^2(\lambda + L)^2} \sum_{m=1}^{q} \left\| \delta m^{n,k+m} \right\|^2 .$$

Remark 2 The above contraction result implies that the composite quantity $m^{n+1,k+m}$, symmetric strain $\boldsymbol{\varepsilon}(\boldsymbol{u}_h^{n+1,k})$, and flux $z_h^{n+1,m+k}$ converge at a geometric rate. Relatively straightforward arguments that include induction in finer time steps, standard mixed method for controlling pressure by flux, Korn's inequality to control the H^1 norm by the L^2 norm of the symmetric strain tensor, imply the convergence of $p_h^{n+1,k+m}, \boldsymbol{u}_h^{n+1,k}$ in L^2 and H^1 norms respectively. The limit equations consist of a coupled system of q finer flow steps at $t_{k+m}, m = 0, \ldots, q$ and a mechanics step at t_{k+q}. The details are spared.

Acknowledgements K. Kumar acknowledges the support of Statoil-UiB Akademia grant that allowed him to travel to ENUMATH conference at Ankara and CSM, ICES, UT Austin for the hospitality where part of the work was completed.

References

1. M.A. Biot, General theory of three-dimensional consolidation. J. Appl. Phys. **12**(2), 155–164 (1941)
2. X. Gai, R.H. Dean, M.F. Wheeler, R. Liu, Coupled geomechanical and reservoir modeling on parallel computers, in *The SPE Reservoir Simulation Symposium*, Houston, 3–5 Feb 2003
3. X. Gai, S. Sun, M.F. Wheeler, H. Klie, A timestepping scheme for coupled reservoir flow and geomechanics on nonmatching grids, in *SPE Annual Technical Conference and Exhibition*, Dallas (2005). SPE97054
4. V. Girault, K. Kumar, M.F. Wheeler, Convergence of iterative coupling of geomechanics with flow in a fractured poroelastic medium. Ices report 15-05, Institute for Computational Engineering and Sciences, The University of Texas at Austin, Austin, 2015
5. J. Kim, H.A. Tchelepi, R. Juanes, Stability, accuracy, and efficiency of sequential methods for coupled flow and geomechanics, in *The SPE Reservoir Simulation Symposium*, Houston, 2–4 Feb 2009. SPE119084
6. A. Mikelić, M.F. Wheeler, Convergence of iterative coupling for coupled flow and geomechanics. Comput. Geosci. **17**, 455–461 (2013)
7. A. Mikelić, B. Wang, M.F. Wheeler, Numerical convergence study of iterative coupling for coupled flow and geomechanics. Comput. Geosci. **18**, 325–341 (2014)
8. P.J. Phillips, M.F. Wheeler, A coupling of mixed and continuous Galerkin finite element methods for poroelasticity. I. The continuous in time case. Comput. Geosci. **11**(2), 131–144 (2007)

9. I.S. Pop, F. Radu, P. Knabner, Mixed finite elements for the Richards' equation: linearization procedure. J. Comput. Appl. Math. **168**(1–2), 365–373 (2004)
10. F.A. Radu, J.M. Nordbotten, I.S. Pop, K. Kumar, A robust linearization scheme for finite volume based discretizations for simulation of two-phase flow in porous media. J. Comput. Appl. Math. **289**, 134–141 (2015)
11. A. Settari, F.M. Mourits, Coupling of geomechanics and reservoir simulation models. in *Computer Methods and Advances in Geomechanics*, ed. by H.J. Siriwardane, M.M. Zema (Balkema, Rotterdam, 1994), pp. 2151–2158
12. L. Shan, H. Zheng, W.J. Layton, A decoupling method with different subdomain time steps for the nonstationary Stokes-Darcy model. Numer. Methods Part. Differ. Equ. **29**(2), 549–583 (2013)
13. R.E. Showalter, Diffusion in poro-elastic media. J. Math. Anal. Appl. **251**(1), 310–340 (2000)
14. M.F. Wheeler, I. Yotov, A multipoint flux mixed finite element method. SIAM J. Numer. Anal. **44**, 2082–2106 (2006)

Part VII
Computational Fluid Dynamics

CFD Simulation of Interaction between a Fluid and a Vibrating Profile

Petr Furmánek and Karel Kozel

Abstract This work deals with numerical simulation of incompressible flow over a profile vibrating with two degrees of freedom. The profile can oscillate around prescribed elastic axis and vibrate in vertical direction and its motion is induced by the flow. The finite volume method was chosen for the solution, namely the so called Modified Causon's Scheme, which is derived from TVD form of the classical predictor-corrector MacCormack scheme and enhanced with the use of the Arbitrary Lagrangian-Eulerian method in order to simulate unsteady flows. Various initial settings are considered (different inlet velocities, initial deviation angles and shifts in vertical direction). Stiffness is modelled both as linear and non-linear. Obtained results are compared with NASTRAN analysis (Čečrdle and Maleček, Verification FEM model of an aircraft construction with two and three degrees of freedom. Technical report R-3418/02, Aeronautical Research and Test Establishment, Prague, Letňany, 2002. In Czech). The resulting critical velocities for unstable oscillations are in the same interval for all simulated cases.

1 Introduction

Aero-elastic effects like buffeting or flutter, which can occur in flows around profiles and wings, have a significant influence on both flow-field and vibrating solid body. However, possibility to simulate them numerically with the use of commercial CFD codes is still very limited and such problems are often solved by a problem-tailored in-house developed software solvers. The authors decided to develop such a CFD in-house code, that could simulate incompressible inviscid low Mach number flow over a profile with flow-induced vibrations by considering 2 degrees of freedom. Unlike in [7], where the problem is solved by the finite element method using unstructured mesh, the authors have chosen the finite volume method and structured computational mesh. The Modified Causon's scheme [6], which is

P. Furmánek (✉) • K. Kozel
Faculty of Mechanical Engineering, Department of Technical Mathematics, CTU in Prague, Karlovo Náměstí 13, Praha 2, 12135, Czech Republic
e-mail: petr.furmanek@fs.cvut.cz; karel.kozel@fs.cvut.cz

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_42

based on TVD version of the classical predictor-corrector MacCormack scheme proposed by Causon [1], was chosen for the solution. The scheme is not TVD, but retains almost the same level of accuracy and saves almost 30 % of computational time and memory. In order to model unsteady effects, the scheme is rewritten in the arbitrary Lagrangian–Eulerian (ALE [3]) form. For testing purposes, the flow around the standard NACA 0012 profile was considered. The elastically supported profile could rotate around the elastic axis and oscillate in the vertical direction. Its motion is induced by the flowing air and described by system of two 2nd order ordinary differential equations. More inlet velocities, initial deviation angles and shifts in the vertical direction are considered. Stiffness is modelled both as linear and non-linear, but for the chosen flow regimes the difference is negligible. Obtained results are compared with NASTRAN analysis [2] and the critical velocities for unstable oscillations are in the same interval for all simulated cases.

2 Mathematical Description of the Problem

This section presents mathematical description of the used models as they were implemented in the solver.

2.1 Governing System of Equations

The investigated problem is considered as two-dimensional, incompressible, inviscid and unsteady and is solved on time interval $[0, T]$ in time-dependent computational domain Ω_t. It is therefore described by the system of incompressible Euler equations, which can be written in the following vector form

$$(\mathbb{D}\mathbf{W})_t + \mathbf{F}_x + \mathbf{G}_y = 0, \tag{1}$$

where

$$\mathbb{D} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \mathbf{W} = \begin{pmatrix} \mathrm{p} \\ u \\ v \end{pmatrix}, \mathbf{F} = \begin{pmatrix} u \\ u^2 + \mathrm{p} \\ uv \end{pmatrix}, \mathbf{G} = \begin{pmatrix} v \\ uv \\ v^2 + \mathrm{p} \end{pmatrix}. \tag{2}$$

W is vector of conservative variables, **F** and **G** are convective fluxes, (u, v) are components of the velocity vector in the directions of the Cartesian coordinates and p is kinematic pressure ($\mathrm{p} = \frac{p}{\rho}$). In order to get better convergence, the system (1) is further rewritten with the use of the artificial compressibility method. Its principle consists of modifying the governing equations by introducing time derivative of pressure into the continuity equation. Hence the first component of the matrix $\mathbb{D}$ is changed as $\mathbb{D}_{11} = \frac{1}{\beta^2}$ (modified matrix is then renamed as $\mathbb{D}_\beta$), where $\beta \in \mathbb{R}^+$ is an artificial compressibility coefficient. In case of unsteady flows β should ideally approach infinity ($\beta \longrightarrow \infty$). For numerical simulations value of $\beta = 10$ was chosen as a reasonable compromise between precision and convergence speed. The resulting system of governing equations with consideration of the artificial compressibility method is then [4]

$$\mathbf{W}_t + (\mathbb{D}_\beta^{-1}\mathbf{F})_x + (\mathbb{D}_\beta^{-1}\mathbf{G})_y = 0. \tag{3}$$

2.2 Description of the Profile Motion

Profile is considered with two degrees of freedom i.e. it can oscillate around elastic axis and vibrate in direction of the vertical axis. Such motion is generally described by the following system of two 2nd order ordinary differential equations

$$\begin{aligned} m\ddot{h} + S_\varphi\ddot{\varphi} + k_{hh}h + d_{hh}\dot{h} &= -L(t), \\ S_\varphi\ddot{h} + I_\varphi\ddot{\varphi} + k_{\varphi\varphi}\varphi + d_{\varphi\varphi}\dot{\varphi} &= M(t). \end{aligned} \tag{4}$$

Here time dependent torsional moment M and lift force L are defined as $M(t) = d\oint_{\Gamma(t)} \mathrm{p}\,\mathbf{r}\cdot\mathbf{n}^{\perp} dl$and $L(t) = d\oint_{\Gamma(t)} \mathrm{p}\, n_2\, dl$ respectively, where Γ is profile boundary, d is profile depth, $\mathbf{r} = (x - x^{EA}, y - y^{EA})$ is position vector of a point on profile surface with respect to the elastic axis (x^{EA}, y^{EA}), $\mathbf{n} = (n_1, n_2)$ is unit inner normal to the profile surface, m is profile mass, S_φ is profile static moment around the elastic axis EA, I_φ is profile inertia moment around the elastic axis EA, k_{hh} and $k_{\varphi\varphi}$ are bending and torsional stiffness of supporting springs, respectively, and d_{hh} and $d_{\varphi\varphi}$ are coefficients of proportional damping (Fig. 1).

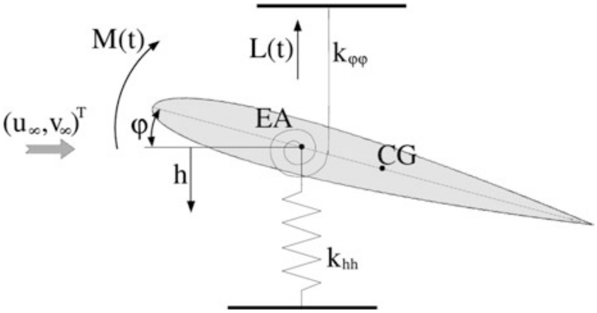


Fig. 1 Profile with 2 degrees of freedom

3 Numerical Solution

Numerical methods for solution of coupled system formed by (3) and (4) are discussed in this section.

3.1 *Modified Causon's Scheme for FVM*

Modified Causon's scheme (MCS) is based on the classical explicit MacCormack predictor-corrector scheme in TVD form, which is able to deliver very good results. However, it also entails disadvantageous demands for computational memory and power. A simplification (based on modification already introduced by Causon [1]) was therefore proposed by the authors [5]. It saves approximately 30 % of computational time but keeps the same level of precision as the original TVD scheme. Due to the use of the ALE method for unsteady flow modelling additional terms appear in both predictor and corrector steps and for their evaluation three mesh configurations have to be used during one time step.

3.2 *Profile Motion Solution*

In order to solve (4) numerically, it is favourable to rewrite it as a system of four 1st order ordinary differential equations. Then, the system is solved by the standard 4th order Runge–Kutta method for ODE's.

3.3 *Mesh Modification for the ALE Computation*

The actual position of mesh vertices during the unsteady computation with the use of the ALE method is given by the following prescription

$$\mathbf{x}(t) = \mathbb{Q}\big[\phi(t, ||\mathbf{x}(0) - \mathbf{x}_{ref}||)\big](\mathbf{x}(0) - \mathbf{x}_{ref}) + \mathbf{x}_{ref} + \mathbf{h}, \tag{5}$$

where

$$\mathbb{Q}(\phi) = \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix}, \ \mathbf{h} = (0, h(t)) \tag{6}$$

and

$$\phi(t,r) = \begin{cases} \varphi(t) & \text{for } r < r_1, \\ \varphi(t) f_D(r) & \text{for } r_1 \le r < r_2, \\ 0 & \text{for } r_2 < r. \end{cases} \tag{7}$$

where

$$f_D(r) = \left[2\left(\frac{r-r_1}{r_2-r_1}\right)^3 - 3\left(\frac{r-r_1}{r_2-r_1}\right)^2 + 1 \right]. \tag{8}$$

It means that the circle centered at $\mathbf{x}_{ref}$ and having radius r_1 is rotating and shifting in vertical direction (according to the values of pitching angle φ and shift h obtained as solution of (4)) as a solid body. The outer area of the second circle with the radius $r_2 > r_1$ is motionless and in the annulus between these two circles the mesh motion is damped by damping function$f_D(\cdot\,)$

3.4 Boundary Conditions

Boundary conditions were set in a standard way for inviscid incompressible flow and realized with the use of *ghost cells* located outside of the computational field.

- At the **inlet** velocity vector was prescribed, so $\mathbf{u_{in}} = (u_\infty, v_\infty)^T$ and pressure p was extrapolated from the flow field.
- At the **outlet** velocity vector $\mathbf{u_{out}}$ was extrapolated from the flow field and outlet pressure p_{out} was prescribed.
- On the **slip wall**, the reflection principle was used for velocity vector, i.e.

$$(u_{ij}^{sw}, v_{ij}^{sw})^T = (u_{ij}, v_{ij})^T - 2\Big[(u_{ij}, v_{ij})^T \cdot \mathbf{n}^{sw}\Big]\mathbf{n}^{sw} + \\ + 2\Big[(u_w, v_w)^T \cdot \mathbf{n}^{sw}\Big]\mathbf{n}^{sw}, \tag{9}$$

where $\mathbf{n}^{sw}$ is normal vector to the given slip wall and indices $_{ij}$ and $_w$ signify velocity in cell i, j in vicinity of the slip wall and wall velocity during the unsteady motion. For pressure boundary condition the *Curvature Corrected Symmetry Technique* (CCST) was chosen. Its basic idea is to use the local momentum equation for pressure evaluation:

$$\mathrm{p}_{ij}^{sw} = \mathrm{p}_{ij} - \triangle n^{sw} \frac{\|(u_w, v_w)^T\|^2}{R}, \tag{10}$$

where $\triangle n^{sw}$ is distance between the centres of gravity of the ghost cell C_{ij}^{sw} and computational cell C_{ij} and R is the radius of the local curvature.

4 Numerical Results

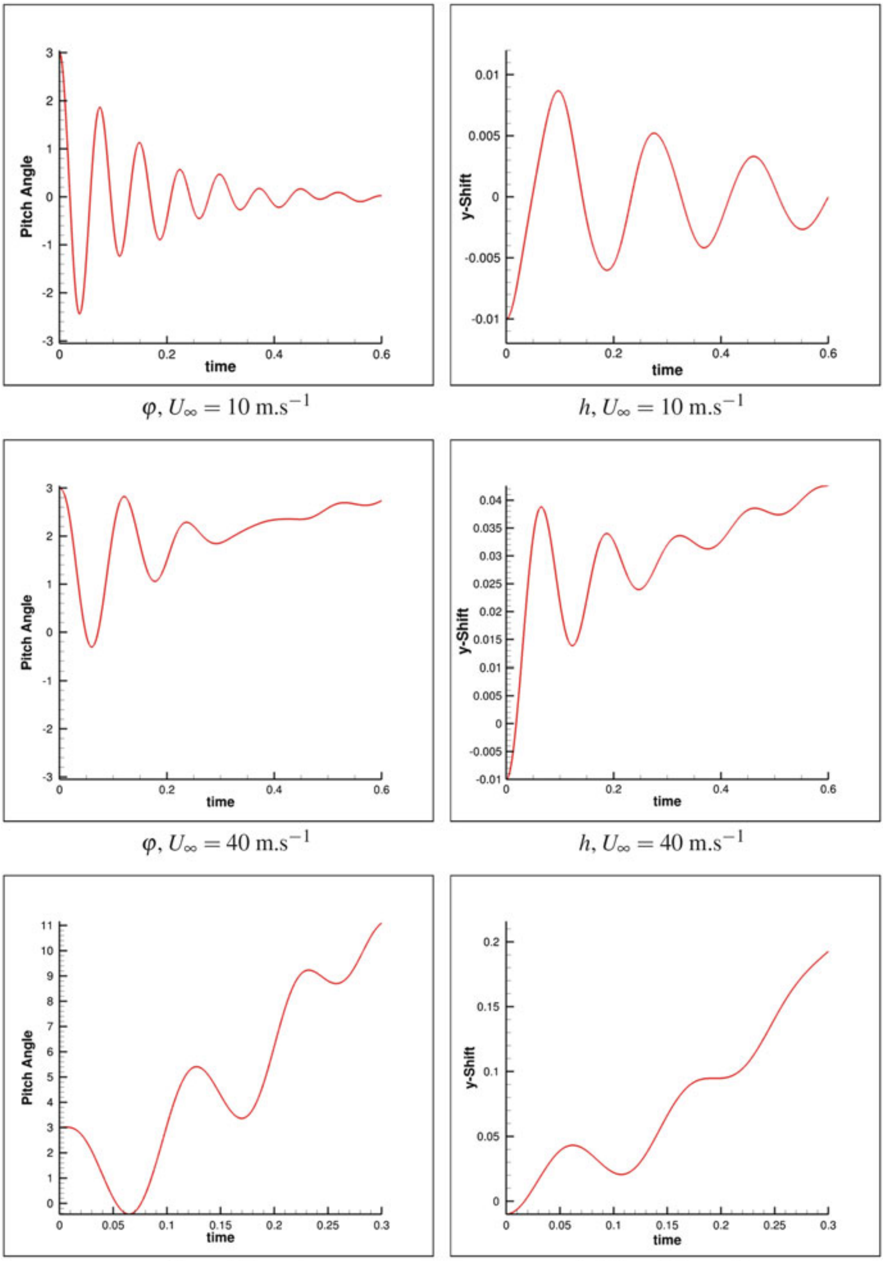
Numerical simulations were carried out for NACA 0012 profile with chord $c = 0.3\,[\mathrm{m}]$, depth $d = 0.05\,[\mathrm{m}]$ and mass $m = 0.086622\,[\mathrm{kg}]$. Inlet velocity was considered in range $U_\infty = 5, 10, 15, 20, 25, 30, 35, 40, 45\,[\mathrm{m.s}^{-1}]$ while standard atmospheric characteristics were prescribed for pressure and density $p = 101325\,[\mathrm{Pa}]$, $\rho = 1.225\,[\mathrm{kg.m}^{-3}]$. Two sets of initial deviation angles and shifts in vertical axis were considered

1. $\varphi_1 = 3^\circ, h_1 = -0.01$ [m],
2. $\varphi_2 = 6^\circ, h_2 = -0.05$ [m].

Aeroelastic properties of the coupled systems were: moment of inertia $I_\varphi = 0.000487291\,[\mathrm{kg.m}^2]$, static moment $S_\varphi = -0.000779673\,[\mathrm{kg.m}]$, torsional stiffness $k_{\varphi\varphi} = 3.695582\,[\mathrm{N.m.rad}^{-1}]$, torsional damping $d_{\varphi\varphi} = 10^{-3} \cdot k_{\varphi\varphi}\,[\mathrm{N.m.rad}^{-1}]$, bending stiffness $k_{hh} = 105.109\,[\mathrm{N.m}^{-1}]$ and bending damping $d_{hh} = 10^{-3} \cdot d_{\varphi\varphi}$ $[\mathrm{N.m}^{-1}]$.

Obtained behaviour of the pitching angle φ (vertical axis of left figures) and vertical shift h (vertical axis of right figures) are plotted on Figs. 2 and 3. Behaviour of the system shows same tendency for both sets of initial conditions. For smaller inlet velocities, the unsteady motion is damped and damping influence of aerodynamic forces increases with increasing inlet velocity so both pitching angle and vertical shift displacements converge to zero.

This is an expected result as the NACA 0012 profile is symmetrical and the inlet velocities are in the sub-critical range. Limiting velocity for stable behaviour is $35\,\mathrm{m.s}^{-1}$. Above this limit (i.e. for velocities about $40\,\mathrm{m.s}^{-1}$) divergence appears and both φ and h increase to relatively large values ($\varphi > 11^\circ$ and $h > 0.2\,\mathrm{m}$). In reference [2] a flutter analysis is performed in NASTRAN with the aid of the strip model for the same flow regimes as are investigated in this work. The critical velocity for divergence is $37.7\,\mathrm{m.s}^{-1}$. Results shown in Figs. 2 and 3 are in good agreement with the NASTRAN simulation.

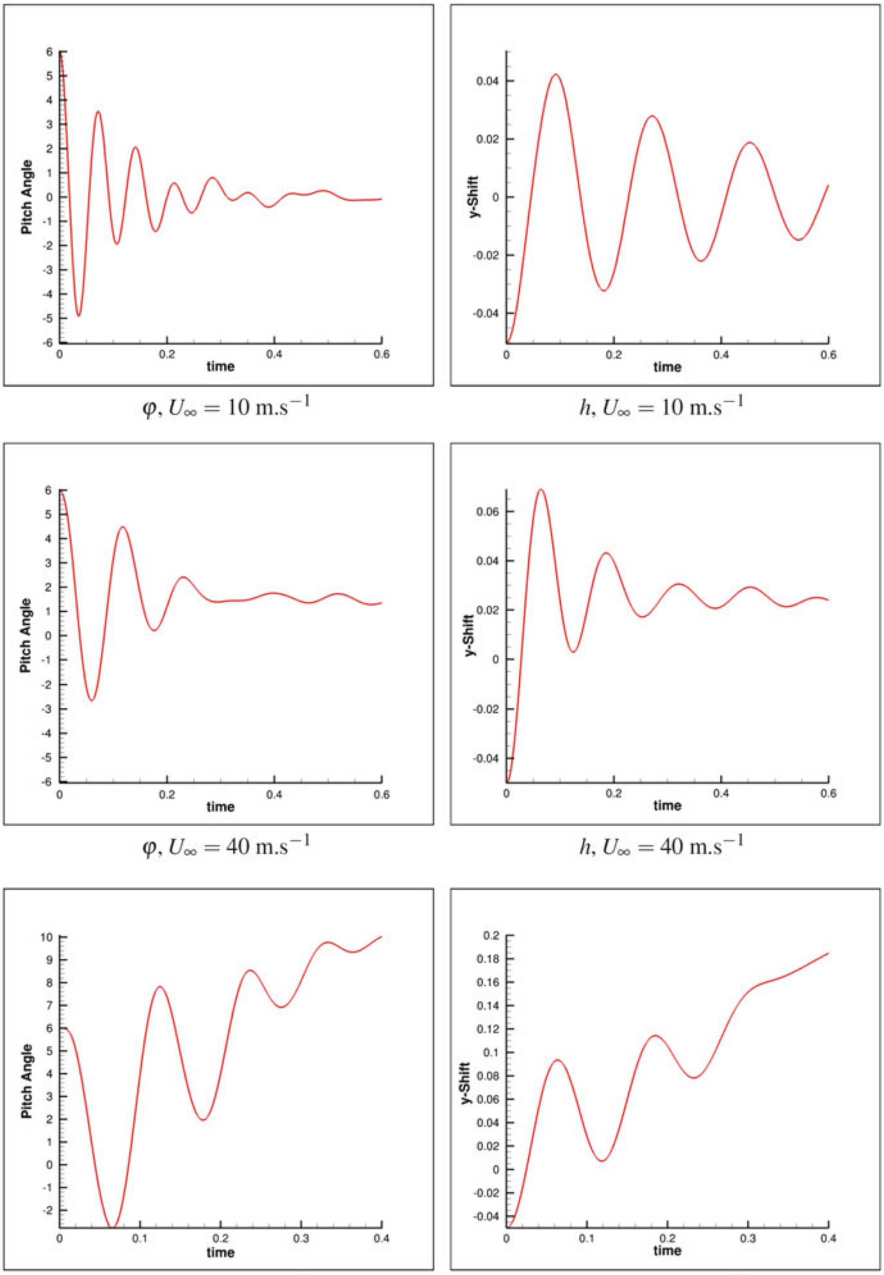


$\varphi, U_\infty = 10\ \mathrm{m.s}^{-1}$ $h, U_\infty = 10\ \mathrm{m.s}^{-1}$

$\varphi, U_\infty = 40\ \mathrm{m.s}^{-1}$ $h, U_\infty = 40\ \mathrm{m.s}^{-1}$

$\varphi, U_\infty = 45\ \mathrm{m.s}^{-1}$ $h, U_\infty = 45\ \mathrm{m.s}^{-1}$

Fig. 2 Initial values: $\varphi_1 = 3°, h_1 = -0.01$ [m], behaviour of pitch angle φ and y-shift h for inlet velocities $U_\infty \in [10, 40, 45]\,\mathrm{m.s}^{-1}$, first 0.6 s (0.3 s resp.) of flow



$\varphi, U_\infty = 10\ \mathrm{m.s}^{-1}$

$h, U_\infty = 10\ \mathrm{m.s}^{-1}$

$\varphi, U_\infty = 40\ \mathrm{m.s}^{-1}$

$h, U_\infty = 40\ \mathrm{m.s}^{-1}$

$\varphi, U_\infty = 45\ \mathrm{m.s}^{-1}$

$h, U_\infty = 45\ \mathrm{m.s}^{-1}$

Fig. 3 Initial values: $\varphi_2 = 6^\circ, h_2 = -0.05$ [m], behaviour of pitch angle φ and y-shift h for inlet velocities $U_\infty \in [10, 40, 45]\ \mathrm{m.s}^{-1}$, first 0.6 s (0.3 s resp.) of flow

5 Conclusion

The paper presents a fast and reliable method for numerical simulation of fluid-structure interaction problems with two degrees of freedom when considering inviscid incompressible flow. The used numerical scheme is based on TVD approach and can be easily extended for compressible problems as well. In the future work, the authors aim for an implementation of implicit version of the scheme and its extension for viscous problems.

Acknowledgements This work was partially supported by grant GACR P101/11/0207.

References

1. D.M. Causon, High resolution finite volume schemes and computational aerodynamics, in *Nonlinear Hyperbolic Equations – Theory, Computation Methods and Applications*, ed. by J. Ballmann, R. Jeltsch. Volume 24 of Notes on Numerical Fluid Mechanics (Vieweg, Braunschweig, 1989), pp. 63–74
2. J. Čečrdle, J. Maleček, Verification FEM model of an aircraft construction with two and three degrees of freedom. Technical report R-3418/02, Aeronautical Research and Test Establishment, Prague, Letňany, 2002. In Czech
3. J. Donea, An arbitrary Lagrangian-Eulerian finite element method for transient fluid-structure interactions. Comput. Methods Appl. Mech. Eng. **33**, 689–723 (1982)
4. M. Feistauer, J. Felcman, I. Straškraba, *Mathematical and Computational Methods for Compressible Flow*. Numerical Mathematics and Scientific Computation (Oxford University Press, New York, 2003). ISBN:0-19-850588-4
5. P. Furmánek, J. Fürst, K. Kozel, High order finite volume schemes for numerical solution of 2D and 3D transonic flows. Kybernetika **45**, 567–579 (2009)
6. J. Fürst, Numerical solution of transonic flow using modern schemes of finite volume method and finite differences, Dissertation thesis (in Czech), ČVUT, Praha, 2001
7. P. Sváček, Numerical modelling of aeroelastic behaviour of an airfoil in viscous incompressible flow. Appl. Math. Comput. **217**(11), 5078–5086 (2011). ISSN:0096-3003

Chebyshev Spectral Collocation Method for Natural Convection Flow of a Micropolar Nanofluid in the Presence of a Magnetic Field

Önder Türk

Abstract The two-dimensional, laminar, unsteady natural convection flow of a micropolar nanofluid (Al_2O_3-water) in a square enclosure under the influence of a magnetic field, is solved numerically using the Chebyshev spectral collocation method (CSCM). The nanofluid is considered as Newtonian and incompressible, and the nanoparticles and water are assumed to be in thermal equilibrium. The governing equations in nondimensional form are given in terms of stream function, vorticity, micrototaion and temperature. The coupled and nonlinear equations are solved iteratively in the time direction, and an implicit backward difference scheme is employed for the time integration. The boundary conditions of vorticity are computed within this iterative process using a CSCM discretization of the stream function equation. The main advantages of CSCM, such as the high accuracy and the ease of implementation, are made used of to obtain solutions for very high values of *Ra* and *Ha*, up to 10^7 and 1000, respectively.

1 Introduction

The micropolar fluid model introduced by Eringen [4] has been a standard phenomena in continuum mechanics, and has led to a reexamination of many classical fluid flow problems. This model introduces a kinematic variable referred as microrotaion to describe the microstructure of the fluid together with the inertial characteristics of the rotating particles. The subject is a very active field of research, and there are many studies which examine the microstructral effects of nanofluids numerically (see e.g. [6–8, 13]).

A considerable effort has been directed to analyze the flow behavior and heat transfer processes in enclosures filled with electrically conducting fluids in the presence of an external magnetic field. One of the prime interests is to study the effects of the interacting buoyancy force and Lorentz force on the flow and

Ö. Türk (✉)
Gebze Technical University, 41400, Gebze-Kocaeli, Turkey
e-mail: onder.turk@yandex.com

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_43

heat transfer. In particular, in a recent work [1], the micropolar flow theory is incorporated to the natural convection of an electrically conducted nanofluid in a square cavity subjected to a magnetic field. The work is an extension of the micropolar model for nanofluids integrating experimental models into the micropolar theory proposed by the same authors in [2]. In both works, the models are employed to investigate the natural convection of nanofluids numerically by using a meshfree point collocation method with a velocity correction scheme, and very good agreements with experimental findings are reported.

In this paper, the Chebyshev spectral collocation method (CSCM) is applied to obtain numerical solutions to natural convection flow of a micropolar nanofluid. The flow takes place in a square enclosure, and is subjected to an externally applied magnetic field. The unknown vorticity boundary values on the walls of the enclosure are approximated by means of a technique based on CSCM. Apart from its novelty, the presented approach turns out to meet several important requirements, such as capturing thin boundary layers for very high characteristic flow parameters. To the best of the author's knowledge, this study is the first application of a spectral collocation method for solving MHD natural convection flow of a micropolar nanofluid, and the first results achieving very high set of values of Rayleigh and Hartmann numbers are reported.

2 Governing Equations

The micropolar theory is incorporated to investigate the two-dimensional, transient, laminar and incompressible natural convection flow an electrically conductive nanofluid, aluminum oxide (Al_2O_3)-water. The flow takes place in a square cavity which is subjected to a horizontally applied uniform magnetic field. Based on the assumptions given in [1, 11], the governing equations in terms of stream function ψ, vorticity w, microrotation N, and temperature T are given nondimensionally as

$$
\begin{aligned}
\nabla^2\psi &= -w, \\
\frac{\partial w}{\partial t} &= (\frac{\mu_{nf}}{\mu_f} + K)\frac{\rho_f}{\rho_{nf}}\nabla^2 w - K\frac{\rho_f}{\rho_{nf}}\nabla^2 N - \frac{\partial w}{\partial x}\frac{\partial \psi}{\partial y} + \frac{\partial w}{\partial y}\frac{\partial \psi}{\partial x} \\
&\quad + \frac{Ra}{Pr}\frac{\rho_f}{\rho_{nf}}\frac{(\rho\beta)_{nf}}{(\rho\beta)_f}\frac{\partial T}{\partial x} + Ha^2\frac{\sigma_{nf}}{\sigma_f}\frac{\rho_f}{\rho_{nf}}\frac{\partial^2\psi}{\partial x^2}, \\
\frac{\partial N}{\partial t} &= (\frac{\mu_{nf}}{\mu_f} + \frac{K}{2})\frac{\rho_f}{\rho_{nf}}\nabla^2 N - 2K\frac{\rho_f}{\rho_{nf}}N \\
&\quad - \frac{\partial N}{\partial x}\frac{\partial \psi}{\partial y} + \frac{\partial N}{\partial y}\frac{\partial \psi}{\partial x} + K\frac{\rho_f}{\rho_{nf}}w, \\
\frac{\partial T}{\partial t} &= \frac{k_{nf}}{k_f}\frac{(\rho C_p)_f}{(\rho C_p)_{nf}}\frac{1}{Pr}\nabla^2 T - \frac{\partial T}{\partial x}\frac{\partial \psi}{\partial y} + \frac{\partial T}{\partial y}\frac{\partial \psi}{\partial x}.
\end{aligned}
\tag{1}
$$

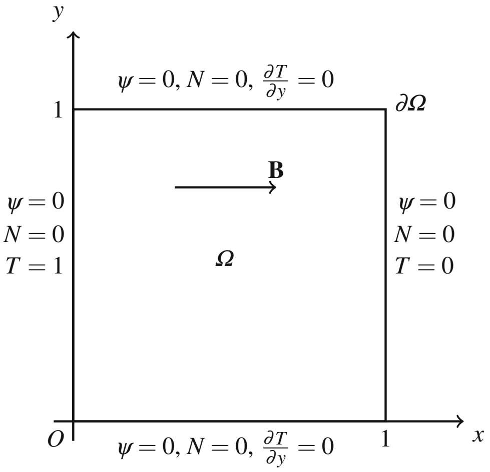

y
$\psi = 0, N = 0, \frac{\partial T}{\partial y} = 0$
1
$\partial\Omega$
B
$\psi = 0$
$N = 0$
$T = 1$
Ω
$\psi = 0$
$N = 0$
$T = 0$
O
$\psi = 0, N = 0, \frac{\partial T}{\partial y} = 0$
1
x

Fig. 1 The problem geometry and the boundary conditions

These equations are supplemented by homogeneous initial conditions and by the boundary conditions illustrated in Fig. 1. The velocity components u and v are related to the stream function and vorticity as $\partial\psi/\partial y = u$, $\partial\psi/\partial x = -v$, and $w = \partial v/\partial x - \partial u/\partial y$. The subscripts '$_{nf}$' and '$_f$' in (2), refer to the nanofluid and base fluid, respectively. K is the material parameter, and Ra, Pr and Ha are the dimensionless Rayleigh, Prandtl and Hartmann numbers, respectively, which are defined as in [1, 11]. The thermo-physical properties of the nanofluid for the present model are determined as in [1, 2]. The electrical conductivity of alumina nanofluid at different volumetric concentrations of alumina nanoparticles is adopted with appropriate units from [12] as $\sigma_{nf} = 0.2983\phi + 0.0058$, and the electrical conductivity of the base fluid is taken accordingly as $\sigma_f = 0.0058$.

3 Numerical Implementation

The spatial discretization of the equations in (2) is based on requiring the numerical approximation of each unknown to be exactly satisfied on the extreme points of the Chebyshev polynomials. Therefore, the technique is referred as Chebyshev spectral collocation method (CSCM). Each function spans the whole domain under consideration, and thus, the derivatives of the function depend on the entire discretization. A function $\Phi(x)$ defined on a certain interval is interpolated by the polynomial $\Phi_L(x)$ of degree at most L of the form $\Phi_L(x) = \sum_{j=0}^{L} C_j(x)\Phi(x_j)$ where $C_j(x)$ is a Cardinal function. The polynomials are differentiated analytically, and a differentiation matrix is constructed for derivative approximation. The higher order derivatives are obtained by multiplying these differentiation matrices, resulting in a high order accurate procedure which is computationally cheap (details can be found in [3, 5, 9, 10]).

The implementation of the unconditionally stable backward difference scheme $\partial u/\partial t\Big|^{m+1} = (u^{m+1} - u^m)/\Delta t$, m and Δt being the time level and the time step, respectively, due to the presence of the temporal derivatives in (2), results in the following CSCM and time discretized matrix-vector equations:

$$\begin{aligned}
&A\psi_L^{m+1} = -w_L^m,\\
&\left[I_L + \Delta t\left(B^{m+1} - (\frac{\mu_{nf}}{\mu_f} + K)\frac{\rho_f}{\rho_{nf}}A\right)\right]w_L^{m+1} =\\
&w_L^m - \Delta tK\frac{\rho_f}{\rho_{nf}}A\,N_L^m + \Delta t\frac{Ra}{Pr}\frac{\rho_f}{\rho_{nf}}\frac{(\rho\beta)_{nf}}{(\rho\beta)_f}F_T^m + \Delta tHa^2\frac{\sigma_{nf}}{\sigma_f}\frac{\rho_f}{\rho_{nf}}F_\psi^{m+1},\\
&\left[I_L + \Delta t\left(B^{m+1} - (\frac{\mu_{nf}}{\mu_f} + \frac{K}{2})\frac{\rho_f}{\rho_{nf}}A + 2K\frac{\rho_f}{\rho_{nf}}\right)\right]N_L^{m+1} =\\
&N_L^m + \Delta tK\frac{\rho_f}{\rho_{nf}}w_L^{m+1},\\
&\left[I_L + \Delta t\left(B^{m+1} - \frac{k_{nf}}{k_f}\frac{(\rho C_p)_f}{(\rho C_p)_{nf}}\frac{1}{Pr}A\right)\right]T_L^{m+1} = T_L^m.
\end{aligned} \tag{2}$$

Here, $(L+1)^2 \times (L+1)^2$ matrices A and B^m are defined as

$$A = I_L \otimes D_L^{(2)} + E_L^{(2)} \otimes I_L,$$

$$B^m = \left(E_L^{(1)} \otimes I_L\right)\psi_L^m\ \left(I_L \otimes D_L^{(1)}\right) - \left(I_L \otimes D_L^{(1)}\right)\psi_L^m\ \left(E_L^{(1)} \otimes I_L\right),$$

where $D_L^{(i)}$ and $E_L^{(i)}$, $i = 1, 2$, are the i-th order Chebyshev differentiation matrices in x and y directions, respectively (see [3, 9] for the details). The $(L+1)^2 \times 1$ vectors F_T and F_ψ are given as

$$F_T^m = \left(I_L \otimes D_L^{(1)}\right)T_L^m, \quad F_\psi^{m+1} = \left(I_L \otimes D_L^{(2)}\right)\psi_L^{m+1},$$

where I_L is the $(L+1)^2 \times (L+1)^2$ identity matrix. In the above equations, $\otimes$ denotes the Kronecker product. The discretized system (2) is solved iteratively. First, the stream function equation is solved by using the initial values of the vorticity. Next, the velocity components are updated by the relation

$$u_L^{m+1} = \left(I_L \otimes D_L^{(1)}\right) \psi_L^{m+1}, \quad v_L^{m+1} = -\left(E_L^{(1)} \otimes I_L\right) \psi_L^{m+1},$$

and, the vorticity boundary values are calculated as

$$\left[w_L^{m+1}\right]\Big|_l = \left[\left(I_L \otimes D_L^{(1)}\right) v_L^{m+1}\right]\Big|_l - \left[\left(E_L^{(1)} \otimes I_L\right) u_L^{m+1}\right]\Big|_l,$$

where l denotes the l-th boundary node. The next step is to obtain the vorticity values on the whole computational domain using the initial values of microrotation and temperature. Finally, the microrotation and temperature equations are solved using the updated stream function and vorticity values. This iterative procedure continues until a preassigned convergence tolerance between two successive iterations is reached for all the unknowns on the whole problem region.

4 Results and Discussions

System (2) is solved by the iterative procedure described in the previous section, setting the convergence tolerance (to steady-state) to be 10^{-5} for all the unknowns. The time step is taken as $\Delta t = 0.001$, and the results are obtained for the values of Prandtl number, material parameter, and relative nanoparticle volumetric fraction, 6.2, 2, and 0.03, respectively. The solutions are obtained by using $L = 50$ for the highest pair of the parameters, namely, $Ra = 10^7$ and $Ha = 1000$, and presented in terms of the contours of the unknowns. The approximations obtained by the application of CSCM are in very good agreement with those obtained using a finite element method scheme in [11]. In particular, Fig. 2 exhibits the good agreement between the CSCM solutions and FEM results for fixed values of $Ra = 10^5$, $Ha = 100$. The same behaviors are observed for all the unknowns, although the contours are drawn on different grids based on the discretization of each method, which is a possible explanation of the slight discrepancies between magnitudes of the contours.

Figures 3, 4, and 5 present the solution contours at a set of Hartmann numbers (0, 100 and 200), for $Ra = 10^4$, 10^5, and 10^6, respectively. Significant variations are observed in the profiles of all the unknowns after the onset of the magnetic field. The suppressing effect of the magnetic force is especially seen on the streamlines and microrotation contours as they tend to have central vortices vertically. These vortices increase in length as Hartmann number increases from 100 to 200. As Ha increases, lower values of stream function and vorticity values in magnitude are observed, indicating a weaker rotation in the central region of the cavity. The

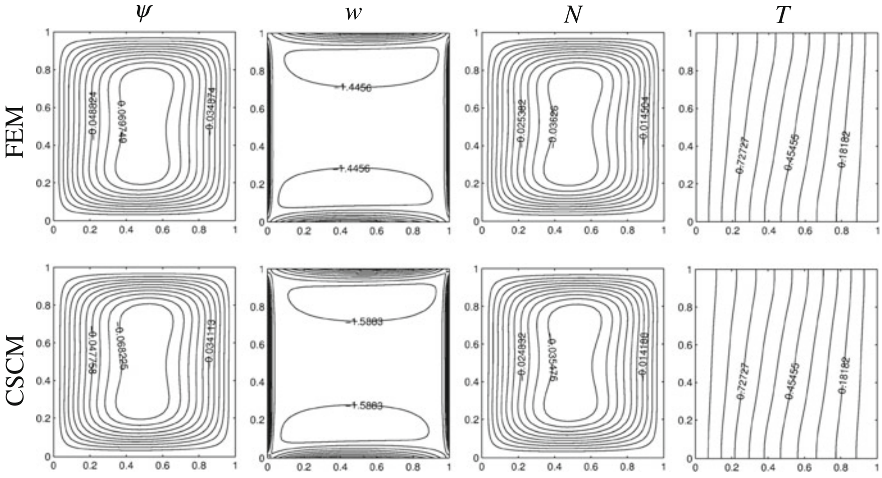


Fig. 2 Comparison of the results obtained by the present scheme with those obtained using FEM where $Ra = 10^5$, $Ha = 100$

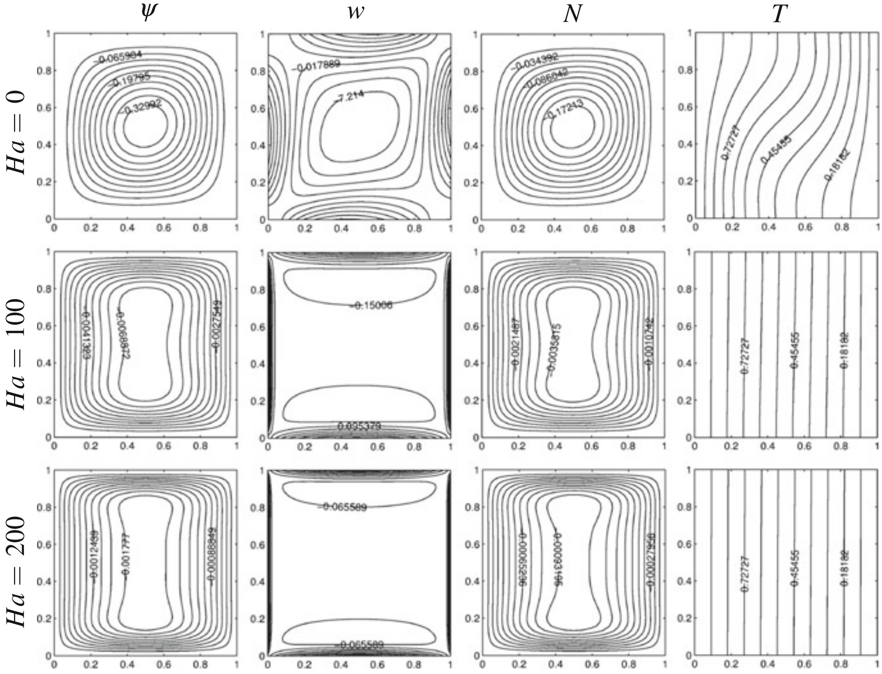


Fig. 3 The effect of Hartmann number when $Ra = 10^4$

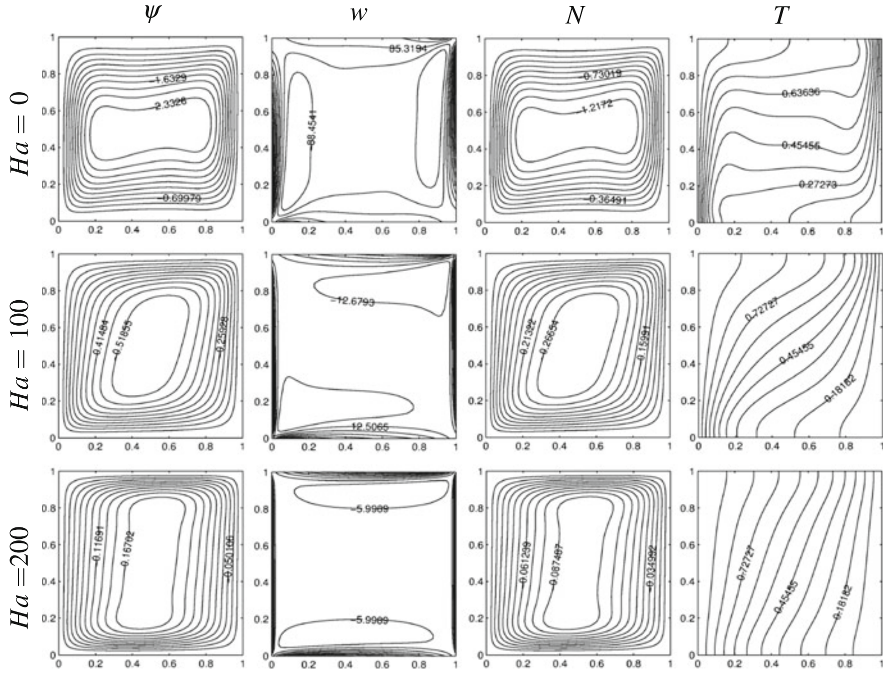


Fig. 4 The effect of Hartmann number when $Ra = 10^6$

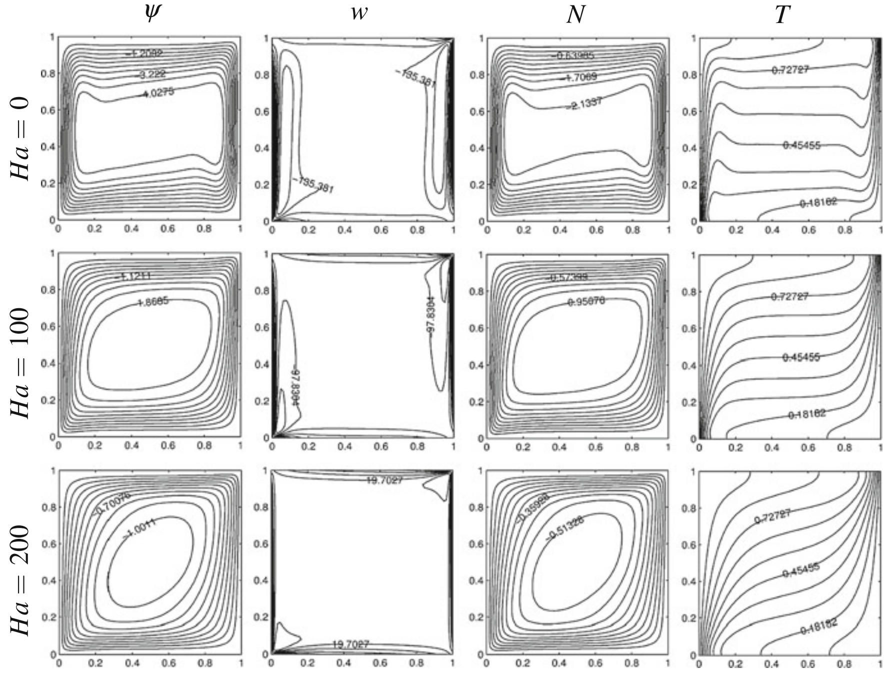


Fig. 5 The effect of Hartmann number when $Ra = 10^7$

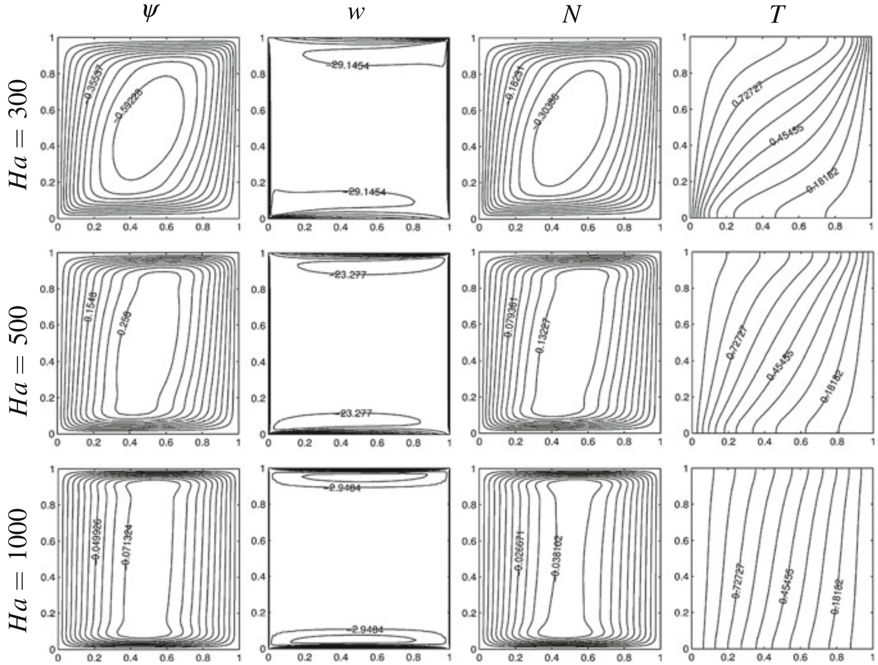


Fig. 6 The effect of Hartmann number when $Ra = 10^7$

buoyancy force lifts the warm fluid particles along the hot wall, and as a result, the isotherms get closer to each other near the hot wall putting forward a higher surface heat flux. For a high value of Rayleigh number, the Lorentz force is evidently suppressed by the buoyant forces, and the effect of the Hartmann number on the flow decreases. The increase in *Ha* results in boundary layers formation in vorticity, and the central region tends to almost stagnant. For a fixed value of $Ra = 10^7$, the effects of increasing *Ha* on the flow and heat transfer can be seen by comparing Fig. 5 (for $Ha = 0$, 100, 200), and Fig. 6 (for $Ha = 300$, 500, 1000). There are formations of condensed boundary layers in the contours of stream function, vorticity and microrotation close to horizontal walls which are parallel to the applied magnetic field. Because of the strong convection, the isotherms are shifted vertically through the adiabatic walls.

5 Conclusion

The simplicity and efficiency of CSCM has been made use of for solving the natural convection flow in enclosures filled with a nanofluid under the effect of a magnetic field. The flow behavior and temperature distribution in the cavity are in excellent agreement with the FEM results. The method presented is shown to be capable of capturing the thin boundary layers, convection or conduction dominance behaviors of the flow and temperature for high *Ra* and *Ha* values. The results indicate that the circulation pattern with a buoyancy driven flow is strongly affected with the variation in *Ra* and *Ha* values. Flattening tendency of the flow due to the strong effect of magnetic field is well observed, in consistence with the previously reported results.

References

1. G. Bourantas, V. Loukopoulos, MHD natural-convection flow in an inclined square enclosure filled with a micropolar-nanofluid. Int. J. Heat Mass Transf. **79**, 930–944 (2014)
2. G. Bourantas, V. Loukopoulos, Modeling the natural convective flow of micropolar nanofluids. Int. J. Heat Mass Transf. **68**(0), 35–41 (2014)
3. J.P. Boyd, *Chebyshev and Fourier Spectral Methods* (Dover, New York, 2000)
4. A.C. Eringen, Simple microfluids. Int. J. Eng. Sci. **2**(2), 205–217 (1964)
5. D. Gottlieb, S.A. Orszag, *Numerical Analysis of Spectral Methods: Theory and Applications*, vol. 26 (SIAM, Philadelphia, 1977)
6. R.U. Haq, S. Nadeem, N.S. Akbar, Z.H. Khan, Buoyancy and radiation effect on stagnation point flow of micropolar nanofluid along a vertically convective stretching surface. IEEE Trans. Nanotechnol. **14**(1), 42–50 (2015)
7. S.T. Hussain, S. Nadeem, R. Ul Haq, Model-based analysis of micropolar nanofluid flow over a stretching surface. Eur. Phys. J. Plus **129**(8), 1–10 (2014)
8. S.K. Jena, S. Bhattacharyya, The effect of microstructure on the thermal convection in a rectangular box of fluid heated from below. Int. J. Eng. Sci. **24**(1), 69–78 (1986)
9. L.N. Trefethen, *Spectral Methods in Matlab* (SIAM, Philadelphia, 2000)
10. Ö. Türk, M. Tezer-Sezgin, Chebyshev spectral collocation method for unsteady MHD flow and heat transfer of a dusty fluid between parallel plates. Numer. Heat Transf. Part A: Appl. **64**(7), 597–610 (2013)
11. Ö. Türk, M. Tezer-Sezgin, Fem solution to natural convection flow of a micropolar nanofluid in the presence of a magnetic field. Meccanica (2015, in press). doi:10.1007/s11012–016–0431–1
12. K.F.V. Wong, T. Kurma, Transport properties of alumina nanofluids. Nanotechnology **19**(34), 345702 (2008)
13. M. Zadravec, M. Hribersek, L. Skerget, Natural convection of micropolar fluid in an enclosure with boundary element method. Eng. Anal. Bound. Elem. **33**(4), 485–492 (2009)

Drag Reduction via Phase Randomization in Turbulent Pipe Flow

Ozan Tugluk and Hakan I. Tarman

Abstract In this study, possibility of reducing drag in turbulent pipe flow via phase randomization is investigated. Phase randomization is a passive drag reduction mechanism, the main idea behind which is, reduction in drag can be obtained via distrupting the wave-like structures present in the flow. To facilitate the investigation flow in a circular cylindrical pipe is simulated numerically. DNS (direct numerical simulation) approach is used with a solenoidal spectral formulation, hence the continuity equation is automatically satisfied (Tugluk and Tarman, Acta Mech 223(5):921–935, 2012). Simulations are performed for flow driven by a constant mass flux, at a bulk Reynolds number (Re) of 4900. Legendre polynomials are used in constructing the solenoidal basis functions employed in the numerical method.

1 Introduction

There are several drag reduction strategies for wall bounded flows, for example, spanwise wall oscillations [2, 4, 6, 12, 15, 16], particle addition [3], and phase randomization [5, 17]. Drag reduction methods can be classified as active or passive, based on whether force is done on the fluid. Phase randomization is a passive drag reduction mechanism. The main idea behind phase randomization is, reduction in drag can be obtained via disrupting the wave-like structures in the flow. This is thought to impede the energy transfer between the wave-like structures and the roll-like structures in the flow field [5]. Application of periodic phase randomization to some of the wave modes is a method proposed to impede this energy transfer, which was first undertaken for the case of channel flow by Sirovich and Handler [5] in a numerical manner and later backed by an experimental study [17]. Both the mentioned works report a drag reduction of up to around 50 %, depending on the

O. Tugluk (✉)
METU Center For Wind Energy (METUWIND/RUZGEM), Ankara, Turkey
e-mail: tugluk@metu.edu.tr

H.I. Tarman
Department of Engineering Sciences, METU, Ankara, Turkey
e-mail: tarman@metu.edu.tr

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_44

mode selection and the disruption frequency. It was also shown that randomizing the phases of high wave number modes or rolls results in a drag increase. This work is built upon our previous work, [18] where the numerical method was developed and tested, and [19] where effects of spanwise wall oscillations on skin friction were investigated. Using the same numerical framework, this study concentrates on the drag reducing effects of phase randomization.

2 Numerical Method

The unsteady incompressible Navier-Stokes equations are non-dimensionalized first. The non-dimensionalization uses the bulk velocity u_B, and the pipe radius R, resulting in Reynolds number $\mathrm{Re} = \frac{2u_B R}{\nu}$, where ν is the kinematic viscosity. The N-S equations and the boundary conditions are then,

$$\begin{aligned} \partial_t \mathbf{u} + (\mathbf{u}\cdot\nabla)\mathbf{u} &= \mathrm{G}(t)\,\mathbf{e_z} - \nabla p + \frac{1}{\mathrm{Re}}\,\Delta\mathbf{u} \\ \nabla\cdot\mathbf{u} &= 0 \\ \mathbf{u}(1,\theta,z,t) = 0, &\quad \mathbf{u}(r,\theta,z,0) = \mathbf{u_0}. \end{aligned} \tag{1}$$

Where $\mathrm{G}(t)$ is the time dependent pressure gradient along the pipe required to obtain constant mass flux. Equation set (1) is solved using solenoidal basis functions for the space discretization [10, 13], and an IMEX (semi-implicit) [1] scheme for time discretization [18, 19]. A similar approach utilizing the solenoidal basis functions for the study of turbulence transition and transient growth of perturbations for pipe flow are detailed in [8, 9] where Chebyshev polynomials are used. In the presented study, Legendre polynomials are used in constructing the solenoidal basis functions, resulting in simpler formulation, as the associated weight is unity. Discretization in space is handled by equidistant grids in the axial and azimuthal (periodic) directions. Along the radial direction a Gauss-Legendre grid is used. Time is discretized in an equispaced fashion using a time step of 5×10^{-3} non-dimensional time units. The computations along the periodic directions are de-aliased using the $3/2$ rule. The resolution is $53\times126\times227$, along radial, azimutal, axial directions respectively. The pipe length Q is taken to be 10 R. The statistics are computed using 300 flowfields equispaced in time with a spacing of 2.5 non-dimensional (bulk) time units, after discarding the initial transients.

2.1 Basis Functions and Projection

There are two sets of conditions to be satisfied by the physical Ψ and the dual Φ basis function families. The first one is the solenoidal condition, i.e. the basis functions should be divergence free,

$$\nabla \cdot \Psi = 0, \quad \nabla \cdot \Phi = 0. \tag{2}$$

It is worthwhile to note that the above formulation for the dual bases is only valid if the dual bases are constructed using Legendre polynomials, which are orthogonal with respect to weight unity. If for example Chebyshev polynomials are used, as in [9], the condition becomes $\nabla \cdot (\omega \ \Phi) = 0$, where $\omega = 1/(\sqrt{1-r^2})$, which further complicates the construction of the solenoidal bases. In addition to the above condition, the physical basis functions are required to satisfy the no-slip boundary conditions. On the other hand, the dual bases are only required the satisfy the condition of the vanishing flux through the walls in order to enable the elimination of the pressure term ∇p from the equation set (1), which is detailed in [11]. Together these additional conditions become,

$$\begin{aligned} \Psi(1, \theta, z) &= 0 \\ \Phi(1, \theta, z) \cdot \mathbf{e}_r &= 0. \end{aligned} \tag{3}$$

In pipe flow, the azimuthal direction is naturally periodic, the flow in the axial direction is taken to be periodic, as is common practice in literature to mimic an infinite pipe. This leads to the use of Fourier representation along the corresponding $\theta,\ z$ coordinates and the basis functions take the following form,

$$\begin{aligned} \Psi_{lnm}(r, \theta, z) &= e^{i(n\theta + 2\pi lz/Q)}\ \mathbf{V}_{lnm}(r) \\ \Phi_{lnm}(r, \theta, z) &= e^{i(n\theta + 2\pi lz/Q)}\ \bar{\mathbf{V}}_{lnm}(r). \end{aligned}$$

The superscripts on the basis functions signify the sufficiency of the two degrees of freedom in representing the three components of a solenoidal velocity field as the continuity equation provides the connection between the components. In the Fourier representation, the three components (V_r, V_θ, V_z) of $\mathbf{V}$ and $\bar{\mathbf{V}}$ are required to satisfy the reduced form of the continuity equation,

$$D_+ V_r + \frac{in}{r}\ V_\theta + ilV_z = 0, \tag{4}$$

where $D_+ = D + \frac{1}{r}$ and $D = d/dr$. The basis functions are then constructed to satisfy Eqs. (3) and (4). The regularity requirement in the vicinity of the pipe center [7, 14] and the use of Gauss-Legendre quadrature (ω_k, r_k) in the numerical evaluation of the inner product integrals,

$$(\bar{\mathbf{V}},\mathbf{V}) = \int_0^1 \bar{\mathbf{V}}(r)^* \cdot \mathbf{V}(r)\, rdr = \frac{1}{2}\int_{-1}^{1} \bar{\mathbf{V}}(r)^* \cdot \mathbf{V}(r)\, rdr$$
$$= \frac{1}{2}\sum_{k=0}^{K} \bar{\mathbf{V}}(r_k)^* \cdot \mathbf{V}(r_k)\, r_k\, \omega_k, \tag{5}$$

impose additional requirements on the basis functions such as the evenness of the integrands. Further information and the forms of the basis functions can be found in [19].

Having constructed the basis functions, the projection procedure is now employed to reduce Eq. (1) to a dynamical system via the solenoidal expansion for velocity (6), and the inner product (5).

$$\mathbf{u}(r,\theta,z,t) = \sum_{l=-L}^{L}\sum_{n=-N}^{N}\sum_{m=0}^{M} a_{lnm}(t)\, e^{i(n\theta+2\pi lz/Q)}\, \mathbf{V}_{lnm}(r) \tag{6}$$

Substituting (6) into (1), and projecting the resulting residual onto the dual bases via the inner product (5), result in the dynamical system for the expansion coefficients $\mathbf{a}_{lnm}$ in the form,

$$\mathbf{A}_{lnm'm}\,\dot{\mathbf{a}}_{lnm} = \mathbf{B}_{lnm'm}\,\mathbf{a}_{lnm} - \mathbf{b}_{lnm'} \tag{7}$$

where,

$$\mathbf{B}_{lnm'm} = (\Psi_{lnm'}, \frac{1}{\mathrm{Re}}\Delta\Phi_{lnm} - G)$$
$$\mathbf{b}_{lnm'} = (\Psi_{lnm'}, (\mathbf{u}\cdot\nabla)\mathbf{u}).$$

The pressure term ∇p is eliminated under the projection, only the forcing term G survives. The system (7) is numerically integrated in time using the 3rd order semi-implicit time-solver, based on backward difference and Adams-Bashford methods,

$$(11\mathbf{A} - 6\Delta t\,\mathbf{B})\,\mathbf{a}^{(k+1)} = \mathbf{A}(18\mathbf{a}^{(k)} - 9\mathbf{a}^{(k-1)} + 2\mathbf{a}^{(k-2)}) \tag{8}$$
$$- \Delta t(18\mathbf{b}^{(k)} - 18\mathbf{b}^{(k-1)} + 6\mathbf{b}^{(k-2)}),$$

for the expansion coefficients $\mathbf{a}$, and they are subsequently used to construct the velocity field (6) and the projection $\mathbf{b}$ for the computation of the next step. This scheme, (8), is not self starting, so for the initial steps we use a Runge-Kutta time integrator of corresponding order.

3 Phase Randomization

Conceptually phase randomization is very simple, considering the expansion for velocity (6), where a_{lnm} are the complex time dependent expansion coefficients. At given intervals a random shift of ϕ_{nm} is introduced, such that,

$$a_{lnm} \rightarrow e^{i\phi_{nm}} a_{lnm}. \tag{9}$$

The operation in (9) can also be seen as a velocity dependent forcing (10), obviously this operation preserves continuity, furthermore it does not change the flow energy. Thus phase randomization does no work on the flow [5].

$$\partial_t \mathbf{u} + (\mathbf{u} \cdot \nabla)\mathbf{u} = -\nabla p + \frac{1}{\mathrm{Re}_{\mathrm{ref}}} \Delta \mathbf{u} + \sum_T F_T(\mathbf{u})\delta(t-T). \tag{10}$$

When phase randomization is applied on a flowfield, mean flow characteristics, such as the average velocity and vorticity profile remain the same. The only visible effect is the rotation or translation of vortices with respect to each other. When applied during the simulation at given intervals however, the effects become visible. In this work, the random phase shifts are applied to a lower band of modes, similar to [5],

$$1 < l < 8 \qquad |n| < 5 \quad \text{corresponding to} \quad \sqrt{k_l^2 + k_n^2} \leq k_{max}/6,$$

where k denotes the wave number and $k_{max} = 75 \times 2\pi/L_z$, 75 being the number of positive Fourier modes selected along the axial direction in this study. The phase randomization is applied at predetermined intervals.

4 Results and Conclusion

For the constant mass flux case, a bulk Reynolds number of 4900 was chosen, to facilitate comparison with the oscillatory wall case. The effect of frequency of phase randomization is shown in Table 1, the optimal drag reduction is observed when phase randomization is applied every 4.4 time units ($875\Delta t$), increasing the

Table 1 Effects of phase randomization frequency

Phase random. freq.	Percent drag reduction
$500\Delta t$	Re-laminarization
$850\Delta t$	Re-laminarization
$875\Delta t$	20
$1000\Delta t$	7
$1250\Delta t$	4

Table 2 Turbulence statistics for controlled and uncontrolled flow in the case of constant mass flux

	No control	Spanwise oscillation	Phase randomization
U_{cl}/U_B	1.31	1.34	1.32
U_B/u_τ	14.08	16.37	15.80
U_{cl}/u_τ	18.45	21.80	20.86
c_f	0.01008	0.0075	0.008012
c_f/c_{f0}	1.000	0.746	0.795
G/G_{nc}	1	0.722	0.78
Re_{cl}	6419	6542	6468
Re_τ	174	142	155

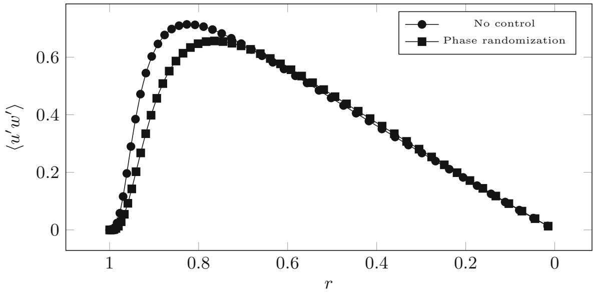


Fig. 1 Mean Reynolds stress for controlled and uncontrolled cases

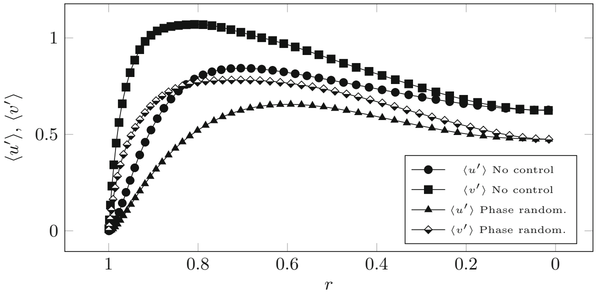


Fig. 2 Root mean square v for controlled and uncontrolled cases

frequency beyond the optimal value results in flow laminarization, where as less frequent application results in lower drag reduction.

The effects on global flow quantities are given in Table 2, together with the effect of spanwise oscillations from an earlier study [19]. It is clear from Table 2 that, phase randomization results in significant reduction (∼20 %) in the required pressure gradient to obtain constant mass-flux (G), and the drag (skin friction, c_f). Also observed is an increase of about 8 % in the mean centerline velocity U_{cl}.

The main effect of phase randomization was found to be transferring turbulent activity away from the wall (Figs. 1 and 2), this was also observed for wall oscillations in [2]. In addition to the observed shift away from the wall, a significant decrease in the magnitude of root mean square velocity components in azimuthal and radial directions was also observed ($15 - 20\,\%$). The reduction in amplitudes, as well as shifting of activity towards the pipe center and away from the wall is evident from Figs. 1 and 2. The maximum reduction in drag was found to be around 20 %, which is lower than achievable by wall oscillations. However, one important point in the comparison is, phase randomization is a passive mechanism and the drag reduction reported here is directly representative of the net power saved, in contrast to wall oscillations, where external energy expenditure is necessary.

References

1. U. Ascher, S.J. Ruuth, B.T.R. Wetton, Implicit-explicit methods for time-dependent pde's. SIAM J. Numer. Anal. **32**, 797–823 (1995)
2. K.S. Choi, M. Graham, Drag reduction of turbulent pipe flows by circular-wall oscillation. Phys. Fluids **10**(1) 7 (1998)
3. J.M.J. Den Toonder, M.A. Hulsen, G.D.C. Kuiken, F.T.M. Nieuwstadt, Drag reduction by polymer additives in a turbulent pipe flow: numerical and laboratory experiments. J. Fluid Mech. **337**, 193–231 (1997)
4. A. Duggleby, K.S. Ball, M.R. Paul, The effect of spanwise wall oscillation on turbulent pipe flow structures resulting in drag reduction. Phys. Fluids **19**(12), 125107 (2007)
5. R.A. Handler, E. Levich, L. Sirovich, Drag reduction in turbulent channel flow by phase randomization. Phys. Fluids A **5**(3), 686–695 (1993)
6. W.J. Jung, N. Mangiavacchi, F.L. Akhavan, Suppression of turbulence in wall-bounded flows by high-frequency spanwise oscillations. Phys. Fluids A **4**(8), 1605–1608 (1992)
7. A. Meseguer, L.N. Trefethen, A spectral Petrov-Galerkin formulation for pipe flow I: Linear Stability and transient growth. Technical report, Oxford University Computing Laboratory, 2000
8. A. Meseguer, L.N. Trefethen, A spectral Petrov-Galerkin formulation for pipe flow: II nonlinear transitional stages. Technical report, Oxford University Computing Laboratory, Oxford, 2001
9. A. Meseguer, L.N. Trefethen, Linearized pipe flow to reynolds number 10^7. J. Comput. Phys. **186**(1), 178–197 (2003)
10. N.M.G. Mhuiris, The construction and use of divergence free vector expansions for incompressible fluid flow calculations. Technical report 86, NASA, 1986
11. R. Moser, P. Moin, A. Leonard, A spectral numerical method for the Navier-Stokes equations with applications to Taylor-Couette flow. J. Comput. Phys. **52**(3), 524–544 (1983)
12. N.V. Nikitin, On the mechanism of turbulence suppression by spanwise surface oscillations. Fluid Dyn. **35**(2), 185–190 (2000)
13. F. Pasquarelli, A. Quarteroni, G. Sacchi-Landriani, Spectral approximations of the Stokes problem by divergence-free functions. J. Sci. Comput. **2**(3), 195–226 (1987)
14. V. Priymak, Accurate Navier-Stokes investigation of transitional and turbulent flows in a circular pipe. J. Comput. Phys. **142**(2), 370–411 (1998)
15. M. Quadrio, P. Ricco, Critical assessment of turbulent drag reduction through spanwise wall oscillations. J. Fluid Mech. **521**, 251–271 (2004)

16. M. Quadrio, S. Sibilla, Numerical simulation of turbulent flow in a pipe oscillating around its axis. J. Fluid Mech. **424**, 217–241 (2000)
17. L. Sirovich, S. Karlsson, Turbulent drag reduction by passive mechanisms. Nature **388**(6644), 728–730 (1997)
18. O. Tugluk, H.I. Tarman, Solenoidal bases for numerical studies of transition in pipe flow. Physica Scripta **T142**, 014009 (2010)
19. O. Tugluk, H.I. Tarman, Direct numerical simulation of pipe flow using a solenoidal spectral method. Acta Mechanica **223**(5), 921–935 (2012)

CFD Optimization of a Vegetation Barrier

Viktor Šíp and Luděk Beneš

Abstract In this study we deal with a problem of particulate matter dispersion modelling in a presence of a vegetation. We present a method to evaluate the efficiency of the barrier and to optimize its parameters. We use a CFD solver based on the RANS equations to model the air flow in a simplified 2D domain containing a vegetation block adjacent to a road, which serves as a source of the pollutant. Modelled physics captures the processes of a gravitational settling of the particles, dry deposition of the particles on the vegetation, turbulence generation by the road traffic and effect of the vegetation on the air flow. To optimize the effectivity of the barrier we employ a gradient based optimization process. The results show that the optimized variant relies mainly on the effect of increased turbulent diffusion by a sparse vegetation and less on the dry deposition of the pollutant on the vegetation.

1 Introduction

Particulate matter (PM) in the atmosphere has a significant negative influence on the human health. It is a concern especially in the urban areas, where the road traffic constitutes a major source of the pollutants. Vegetation barriers were proposed as a means to the reduction of a harmful PM in the atmosphere. Due to the complexity of the problem, assessment of the effectivity of the barriers and its design is difficult without the computer simulations.

Many publications on the topic of mathematical modelling of the pollutant deposition on the vegetation are available. Among the most notable are the following: review [11] on the topic of dry deposition on the vegetation, reviews [5, 9] on the vegetation in urban areas or modelling studies [13, 16, 18].

In this paper we present a method for the evaluation of the effectivity of the barriers and for the numerical optimization of the barrier properties. The model

V. Šíp (✉) • L. Beneš
Faculty of Mechanical Engineering, Institute of Technical Mathematics, Czech Technical University in Prague, 121 35 Prague 2, Karlovo nám. 13, Czech Republic
e-mail: viktor.sip@fs.cvut.cz; ludek.benes@fs.cvut.cz

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_45

presented here is based on the work [15], where the influence of the atmospheric conditions on the barrier efficiency was investigated.

2 Numerical Model

2.1 Physical Model

Let us summarize the basic characteristics of the problem. We are interested in the air flow in the bottom layer of the atmosphere, approximately 200 m thick. Such flow can be modelled as incompressible, but with variable density due to the acting of the gravity force. Three effects of the vegetation should be considered: effect on the air flow, i.e. slowdown or deflection of the flow, influence on the turbulence levels inside and near the vegetation, and the filtering of the particles present in the flow.

2.1.1 Fluid Flow

In our formulation of Reynolds-averaged Navier-Stokes (RANS) equations the pressure p and potential temperature θ are split into background component in hydrostatic balance and fluctuations, $p = p_0 + p'$ and $\theta = \theta_0 + \theta'$. Boussinesq approximation stating that changes in density are negligible everywhere except in the gravity term is utilized. Resulting set of equations is as follows:

$$\nabla \cdot \boldsymbol{u} = 0, \tag{1}$$

$$\frac{\partial \boldsymbol{u}}{\partial t} + (\boldsymbol{u} \cdot \nabla)\boldsymbol{u} + \nabla(p'/\rho_{\text{ground}}) = \nu_E \nabla^2 \boldsymbol{u} + \boldsymbol{g} + \boldsymbol{S_u}, \tag{2}$$

$$\frac{\partial \theta}{\partial t} + \nabla \cdot (\theta \boldsymbol{u}) = \frac{\nu_E}{\text{Pr}} (\nabla \cdot (\nabla \theta)) . \tag{3}$$

Here vector $\boldsymbol{u}$ stands for velocity, ρ_{ground} is the value of the air density ρ at the ground level, $\nu_E = \nu_L + \nu_T$ is the effective kinematic viscosity, which is a sum of the laminar and turbulent viscosity, $\boldsymbol{g} = (0, g\frac{\theta'}{\theta_0}, 0)$ is the gravity term, $\boldsymbol{S_u}$ represent the momentum sink due to the vegetation and $\text{Pr} = 0.75$ is the Prandtl number.

2.1.2 Turbulence

Standard $k - \epsilon$ model is employed to model the turbulence. Equations for turbulence kinetic energy k and dissipation ϵ are as follows:

$$\frac{\partial \rho k}{\partial t} + \nabla \cdot (\rho k \boldsymbol{u}) = \nabla \cdot \left(\left(\mu_L + \frac{\mu_T}{\sigma_k} \right) \nabla k \right) + P_k - \rho\epsilon + \rho S_k, \tag{4}$$

$$\frac{\partial \rho \epsilon}{\partial t} + \nabla \cdot (\rho \epsilon \boldsymbol{u}) = \nabla \cdot \left(\left(\mu_L + \frac{\mu_T}{\sigma_\epsilon} \right) \nabla \epsilon \right) + C_{\epsilon_1} \frac{\epsilon}{k} P_k - C_{\epsilon_2} \rho \frac{\epsilon^2}{k} + \rho S_\epsilon . \tag{5}$$

The model is completed by a relation between k, ϵ and the turbulent dynamic viscosity μ_T, $\mu_T = C_\mu \rho \frac{k^2}{\epsilon}$. In the equations above μ_L is the laminar dynamic viscosity, P_k is the production of the turbulence kinetic energy, and S_k and S_ϵ are sources of k and ϵ respectively. Both consist of a part due to the road traffic and a part due to the vegetation, $S_k = S_k^r + S_k^v$, $S_\epsilon = S_\epsilon^r + S_\epsilon^v$. Sources due to the road traffic are modelled by the model from [2], while sinks and sources due to the vegetation are described below.

Following constants of the model are used: $\sigma_k = 1.0$, $\sigma_\epsilon = 1.167$, $C_{\epsilon_1} = 1.44$, $C_{\epsilon_2} = 1.92$ and $C_\mu = 0.09$.

2.1.3 Particle Transport

Non dimensional mass fraction w of the pollutant in the air is calculated using the equation for the pollutant density,

$$\frac{\partial \rho w}{\partial t} + \nabla \cdot (\rho w \boldsymbol{u}) - (\rho w u_s)_y = \nabla \cdot \left(\frac{\nu_E}{\mathrm{Sc}} \nabla \rho w \right) + \rho f_c + S_w. \tag{6}$$

Here f_c is the source term and S_w is the vegetation deposition term. Based on the review and the discussion in [19], the Schmidt number $\mathrm{Sc} = 0.72$ was used. The settling velocity u_s of a spherical particle with the diameter d and density ρ_p is given by the Stokes' equation, $u_s = (d^2 \rho_p g C_c)/(18\mu)$, with the correction factor $C_c = 1 + \frac{\lambda}{d}(2.34 + 1.05 \exp(-0.39 d/\lambda))$, where $\lambda = 0.066$ μm is the mean free path of the particle in the air [4].

2.1.4 Vegetation

We model the vegetation as horizontally homogenous, described by vertical *Leaf area density* (LAD) profile – foliage surface area per unit volume – and a leaf type (broadleaf or needle) and size of the leaf. Three effects of the vegetation are modelled: first, it is a momentum sink inside the vegetation block, $\boldsymbol{S}_u = -C_d \mathrm{LAD} |\boldsymbol{u}| \boldsymbol{u}$, present in the Eq. (2). Here $C_d = 0.3$ is the drag coefficient [7].

Secondly, it is the influence on the turbulence levels. Following [7], we model this term as

$$S_k^v = C_d \mathrm{LAD}(\beta_p |\boldsymbol{u}|^3 - \beta_d |\boldsymbol{u}| k), \quad S_\epsilon^v = C_{\epsilon 4} \frac{\epsilon}{k} S_k^v,$$

in Eqs. (4) and (5). Constants used are $\beta_p = 1.0$, $\beta_d = 5.1$ and $C_{\epsilon_4} = 0.9$.

And lastly, it is a particle sink term in Eq. (6), $S_w = -\mathrm{LAD} u_d \rho w$. The term is proportional to the deposition velocity u_d. Deposition velocity reflects four main processes by which particles depose on the leaves: Brownian diffusion, interception, impaction and gravitational settling. Its value generally depends on wind speed, particle size and vegetation properties. In this study we adopted the model from [12] derived for broadleaf trees.

2.2 Numerical Methods

2.2.1 CFD Solver

Apart from the divergence constraint (1), all presented PDEs are in a form of a evolution equation, suitable for the discretization as described below. The divergence constraint is transformed into such form by employing method of artificial compressibility with parameter β so that we obtain

$$\frac{1}{\beta}\frac{\partial p'}{\partial t} + \nabla \cdot \boldsymbol{u} = 0. \tag{7}$$

The choice of the parameter β is discussed e.g. in [10], here we have used $\beta = 1000$.

The resulting set of equations is discretized using the finite volume method on unstructured grid. For the convective terms the AUSM + up scheme [8], designed for all speed flows, is used. Second order accuracy is achieved via the linear reconstruction, where gradients are calculated using least squares approach. To prevent artificial overshooting, Venkatakrishan limiter [20] is utilized.

Gradients on the cell faces needed for the calculation of the diffusive terms are evaluated using the Gauss-Green theorem on a dual cell associated with the face.

The discretized system forms a set of ordinary differential equations, which are solved using an implicit BDF2 method. In every time step (outer iteration), first the system of the Navier-Stokes equations (2, 3, 7) is solved, followed by the system of the $k - \epsilon$ equations (4, 5) and then by the system of the passive scalar equations (6). Values of turbulent viscosity, coupling together turbulence equations with the Navier-Stokes equations, are taken from the previous time step.

Each of these nonlinear systems is solved by the Newton method. Inner linear systems are solved using matrix-free GMRES solver. The linear systems are preconditioned by ILU(3) preconditioner. Necessary evaluations of the Jacobians are done via finite differences. Significant cost of these operations is reduced by two complementing approaches: via matrix coloring, which exploit the sparseness of the Jacobian, and by calculating the preconditioner matrices (as well as the Jacobians) only every 20th time step.

Since we are solving only for a steady-state solution, we continuously adapt the time step in order to accelerate the convergence. The adapting criterion is based on the number of the iterations of the linear solvers in one outer iteration. Time stepping proceeds until a steady-state solution is reached.

The solver is written in C++. PETSc library [1] is used for the nonlinear system solution.

2.2.2 Optimization

PDE-constrained optimization problem could be written in the following form:

$$\text{Find } \min_{p \in P} J(\boldsymbol{W}, p) \quad \text{subject to} \quad F(\boldsymbol{W}, p) = 0 \tag{8}$$

and constrained by

$$p_i^{min} \leq p_i \leq p_i^{max} \quad i = 1 \ldots n, \tag{9}$$

$$g_j(p) \leq 0 \quad j = 1 \ldots m. \tag{10}$$

Here $J(\boldsymbol{W}, p)$ is a cost function and $F(\boldsymbol{W}, p)$ is the system of steady-state PDEs, $\boldsymbol{W}$ is the state vector and p is the vector of parameters. Allowed values of parameters are limited by p_i^{min} and p_i^{max}, while functions g_j represents nonlinear constraints.

To solve the optimization problem, method of moving asymptotes (MMA) [17] implemented in NLopt optimization package [6] was employed.

Since the MMA is a gradient-based method, the CFD solver has to facilitate the evaluation of not only the cost function at a given point in the parameter space, but also its derivatives with respect to the parameters. This was done via a direct sensitivity approach [3].

3 Application to the Model Problem

3.1 Case Settings

Figure 1 shows the sketch of the computational domain. Four sources of pollutant, representing the road, are placed between 23 m and 42 m from the inlet at height 0.8 m. Vegetation block of height 15 m is placed downstream from the road.

We model the particles of diameter 10 μm and density 1000 kg/m^3. Each source of the pollutant has the intensity 1 μg/s. No resuspension of the particles fallen on the ground is allowed. Density of the traffic is set to 4 passenger cars and 1 heavy duty vehicle per minute in each of the four lanes.

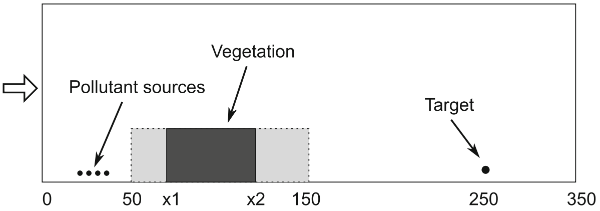


Fig. 1 Sketch of the domain (not to scale)

As in [15], logarithmic wind profile is prescribed at the inlet with $u_{\mathrm{ref}} = 5\,\mathrm{m/s}$ at height $y_{\mathrm{ref}} = 10\,\mathrm{m}$. Roughness parameter z_0 is set to 0.1 m. The atmosphere is under weakly stable stratification ($\partial T/\partial y = 0\,\mathrm{K/m}$). For further details on the boundary conditions for the fluid flow and the pollutant equations see [15]. For the turbulence equations, boundary conditions and wall functions according to [14] are used.

The optimization cost function J is the value of the pollutant concentration at $x = 250\,\mathrm{m}$ from the inlet at height 2 m. Vector of parameters $p = (x_1, x_2, \mathrm{LAI})$ consists of starting and end point of the vegetation block and its *Leaf Area Index*, which is a ratio of a total leaf area relative to the ground area. Following constraints are placed on the parameters:

- Position of the vegetation: $x_{min} \le x_1 \le x_2 \le x_{max}$ with $x_{min} = 50\,\mathrm{m}$ and $x_{max} = 150.0\,\mathrm{m}$.
- Maximal leaf area index: $0.0 \le \mathrm{LAI} \le \mathrm{LAI}_{max}$ with $\mathrm{LAI}_{max} = 9.0$.
- Maximal total amount of trees planted: $(x_2 - x_1)\mathrm{LAI} \le \mathrm{VEG}_{max}$ with $\mathrm{VEG}_{max} = 270.0$. That could represent eg. forest of length 30 m and LAI 9 or length 100 m and LAI 2.7.

3.2 Results

Since our method searches only for a local minimum, three different initial points were used to rule out a possibility that only a local minimum in the vicinity of a initial position was found. The optimization procedure ended in the same point for all of the initial points. The initial configurations and corresponding solutions are listed in Table 1. The optimized variant represents a sparse vegetation block spanning the whole allowed interval. The obtained $\mathrm{LAI} = 0.81$ lies well below the value given by the constraint on the maximal amount of trees planted, which allowed for a $\mathrm{LAI} = 2.7$ for a block spanning the whole interval.

As evident from the Table 1, the cost function (i.e. the concentration behind the barrier) was reduced by 15 %–20 % in all three cases. This reduction is further visible on the left panel of Fig. 2, where the vertical profiles of the particle concentration at $x = 250\,\mathrm{m}$ is shown. Three initial variants and the final variant are complemented by a variant with no vegetation present.

Table 1 Three initial variants and corresponding solutions. The initial and final points are listed in the form of the parameter vector $p = (x_1, x_2, \mathrm{LAI})$

Variant	Initial point	Solution	J (Initial)	J (Final)	#Evaluations
A	(90.0, 110.0, 4.5)	(50.0, 150.0, 0.810)	0.0407	0.0338	39
B	(80.0, 110.0, 6.75)	(50.0, 150.0, 0.810)	0.0419	0.0338	45
C	(60.0, 90.0, 8.1)	(50.0, 150.0, 0.810)	0.0402	0.0338	67

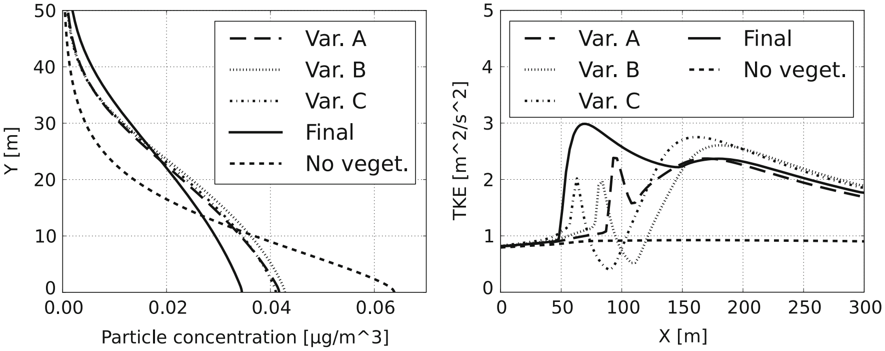


Fig. 2 Vertical profile of particle concentration at x = 250 m (*left*) and horizontal profile of turbulence kinetic energy at height 10 m (*right*)

Table 2 Percentage of the injected pollutant deposed on the vegetation and on the ground

	Variant A (%)	Variant B (%)	Variant C (%)	Final variant (%)
Deposition on the vegetation	2.88	4.30	5.51	2.43
Deposition on the ground	2.88	2.95	2.75	2.32

Table 2 shows that less than 10 % of the injected pollutant was deposed either on the ground or on the vegetation in all cases, and less than 5 % in the optimized variant. The rest was redistributed to the higher layers of the atmosphere, where the higher velocity of the flow allowed for faster dilution. Therefore, the most important effect of the sparse vegetation here is the disturbance of the flow, leading to the increased levels of turbulence and increased turbulent diffusion, which results in faster redistribution to the higher layers. This is demonstrated on the right panel of Fig. 2, where the horizontal profiles of the turbulence kinetic energy are shown for all variants.

4 Discussion

A method for evaluation the effects of vegetation barriers on pollution dispersion was developed and its usability was demonstrated on a simple test case. There are several shortcomings of the method. First, it is suitable only for a limited number of parameters. In the current implementation when 100 parameters are optimized the amount of time for the CFD solution in every step of the optimization loop is roughly equal to the time needed for the gradient evaluation. For higher number of parameters it would be therefore more suitable to use the adjoint method for the gradient calculation.

Secondly, there is a significant uncertainty in vegetation properties, as these are difficult to estimate. Quantification of this uncertainty should therefore be in order.

Thirdly, our method optimizes only for a single target, while in reality we may be interested in several targets at once. To take that into account, multi-objective optimization should be employed.

Lastly, optimization procedure sought only for the local minimum. Here we have used multiple initial points to assess whether we have found the global minimum, however, such approach is not sufficiently rigorous and could be difficult to apply when higher number of parameters is used.

Acknowledgements This work was supported by the grant SGS13/174/OHK2/3T/12 of the Czech Technical University in Prague and by the grant TD020357 of the Technology Agency of the Czech Republic.

References

1. S. Balay, S. Abhyankar, M. Adams, J. Brown, P. Brune, K. Buschelman, L. Dalcin, V. Eijkhout, W. Gropp, D. Kaushik, M. Knepley, L. McInnes, K. Rupp, B. Smith, S. Zampini, H. Zhang, PETSc Web page. http://www.mcs.anl.gov/petsc, 2015
2. D. Bäumer, B. Vogel, F. Fiedler, A new parameterisation of motorway-induced turbulence and its application in a numerical model. Atmos. Environ. **39**(31), 5750–5759 (2005)
3. M. Gunzburger, *Perspectives in Flow Control and Optimization* (Society for Industrial and Applied Mathematics, Philadelphia, 2003)
4. W. Hinds, *Aerosol Technology: Properties, Behavior, and Measurement of Airborne Particles* (Wiley, New York, 1999)
5. S. Janhäll, Review on urban vegetation and particle air pollution – deposition and dispersion. Atmos. Environ. **105**, 130–137 (2015)
6. S. Johnson, The NLopt nonlinear-optimization package. http://ab-initio.mit.edu/wiki/index.php/NLopt, 2015
7. G. Katul, L. Mahrt, D. Poggi, C. Sanz, One- and two-equation models for canopy turbulence. Bound. Layer Meteor. **113**, 81–109 (2004)
8. M.-S. Liou, A sequel to AUSM, part II: AUSM+-up for all speeds. J. Comput. Phys. **214**, 137–170 (2006)
9. T. Litschke, W. Kuttler, On the reduction of urban particle concentration by vegetation – a review. Meteorol. Z. **17**, 229–240 (2008)
10. F. Muldoon, S. Acharya, A modification of the artificial compressibility algorithm with improved convergence characteristics. Int. J. Numer. Methods Fluids **55**(4), 307–345 (2007)
11. A. Petroff, A. Mailliat, M. Amielh, F. Anselmet, Aerosol dry deposition on vegetative canopies. Part I: review of present knowledge. Atmos. Environ. **42**, 3625–3653 (2008)
12. A. Petroff, L. Zhang, S. Pryor, Y. Belot, An extended dry deposition model for aerosols onto broadleaf canopies. J. Aerosol Sci. **40**(3), 218–240 (2009)
13. M. Raupach, N. Woods, G. Dorr, J. Leys, H. Cleugh, The entrapment of particles by windbreaks. Atmos. Environ. **35**, 3373–3383 (2001)
14. P. Richards, R. Hoxey, Appropriate boundary conditions for computational wind engineering models using the k-ϵ turbulence model. J. Wind Eng. Ind. Aerodyn. **46&47**, 145–153 (1993)
15. V. Šíp, L. Beneš, Modelling the effects of a vegetation barrier on road dust dispersion. Appl. Mech. Mater. **821**, 105–112 (2016)
16. J. Steffens, Y. Wang, K. Zhang, Exploration of effects of a vegetation barrier on particle size distributions in a near-road environment. Atmos. Environ. **50**, 120–128 (2012)

17. K. Svanberg, A class of globally convergent optimization methods based on conservative convex separable approximations. SIAM J. Optim. **12**(2), 555–573 (2002)
18. A. Tiwary, H. Morvanb, J. Colls, Modelling the size-dependent collection efficiency of hedgerows for ambient aerosols. Aerosol Sci. **37**, 990–1015 (2005)
19. Y. Tominaga, T. Stathopoulos, Turbulent Schmidt numbers for CFD analysis with various types of flowfield. Atmos. Environ. **41**, 8091–8099 (2007)
20. V. Venkatakrishnan, Convergence to steady state solutions of the euler equations on unstructured grids with limiters. J. Comput. Phys. **118**, 120–130 (1995)

Modified Newton Solver for Yield Stress Fluids

Saptarshi Mandal, Abderrahim Ouazzi, and Stefan Turek

Abstract The aim of this contribution is to present a new Newton-type solver for yield stress fluids, for instance for viscoplastic Bingham fluids. In contrast to standard globally defined ('outer') damping strategies, we apply weighting strategies for the different parts inside of the resulting Jacobian matrices (after discretizing with FEM), taking into account the special properties of the partial operators which arise due to the differentiation of the corresponding nonlinear viscosity function. Moreover, we shortly discuss the corresponding extension to fluids with a pressure-dependent yield stress which are quite common for modelling granular material. From a numerical point of view, the presented method can be seen as a generalized Newton approach for non-smooth problems.

1 Introduction

Continuum theory for slow viscoplastic fluids based on corresponding flow rules typically relates the shear stress and the strain rate in a plastic frictional system via Bingham-like constitutive laws

$$\begin{cases} \boldsymbol{\tau} = 2\nu \mathbf{D}(\boldsymbol{u}) + \tau_s \dfrac{\mathbf{D}(\boldsymbol{u})}{\|\mathbf{D}(\boldsymbol{u})\|} & \text{if } \|\mathbf{D}(\boldsymbol{u})\| \neq 0 \\ \|\boldsymbol{\tau}\| \leq \tau_s & \text{if } \|\mathbf{D}(\boldsymbol{u})\| = 0 \end{cases} \tag{1}$$

where $\mathbf{D}(\boldsymbol{u}) = \frac{1}{2}(\nabla \boldsymbol{u} + (\nabla \boldsymbol{u})^T)$ denotes the strain rate tensor, and τ_s denotes the yield stress. The shear stress has two contributions: a viscous part, and a strain rate independent part. Furthermore, for the deformation of dense granular material, the stress and strain rate tensors are always coaxial. So, for unequal stresses, Schaeffer [4] postulated that the stresses contract in the directions of greater stress and expand

S. Mandal • A. Ouazzi • S. Turek (✉)
Institut für Angewandte Mathematik, LSIII, TU Dortmund, Vogelpothsweg 87, D-44227 Dortmund, Germany
e-mail: saptarshi.mandal@math.tu-dortmund.de; abderrahim.ouazzi@math.tu-dortmund.de; ture@featflow.de

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_46

in directions of smaller stress. As a consequence, the deviatoric part of the related Schaeffer model for flow of dry powder in the quasi-static regime [4] is

$$\boldsymbol{\tau} = \sin(\phi)\, p \frac{\mathbf{D}(\boldsymbol{u})}{\|\mathbf{D}(\boldsymbol{u})\|} \tag{2}$$

where ϕ denotes the angle of internal friction: Hence, this model can be interpreted as pressure-dependent yield stress fluid. Moreover, the interesting transition from solid-like to fluid-like behavior of granular material was investigated experimentally and numerically in [3]. Here, the unified constitutive model for the static and intermediate regimes is given by the following constitutive law (with an appropriate $n > 0$ and $b \in \mathbb{R}^+$, see [3]):

$$\boldsymbol{\tau} = p \left\{ \sin(\phi) + b\cos(\phi)\|\mathbf{D}(\boldsymbol{u})\|^n \right\} \frac{\mathbf{D}(\boldsymbol{u})}{\|\mathbf{D}(\boldsymbol{u})\|} \tag{3}$$

Similarly, Pouliquen et al. [2] proposed an extended constitutive model for dense granular material, where the stress tensor is given as a function of the inertia number $\mathbf{I} = \mathbf{D}(\boldsymbol{u})d/\sqrt{p\rho_p}$ (again with appropriate values of ρ_p and d, see [2])

$$\boldsymbol{\tau} = p\mu(\mathbf{I}) \frac{\mathbf{D}(\boldsymbol{u})}{\|\mathbf{D}(\boldsymbol{u})\|} \tag{4}$$

where $\mu(\cdot)$ is an empirical friction law:

$$\mu(\mathbf{I}) = \mu_1 + \frac{\mu_2 - \mu_1}{\mathbf{I}_0/\mathbf{I} + 1} \tag{5}$$

All models show the relationship between granular and Bingham fluids. In order to incorporate friction into viscoplasticity in mixing wet granular materials, El Khouja et al. [1] introduced the dependency of the pressure in yield stress flow model, i.e. the yield stress $\tau_s(\cdot)$ is a function of the pressure, namely let $\tau_{\min}, \tau_{\max} \in \mathbb{R}^+$, so that $\tau_s(\cdot)$ can be defined as:

$$\tau_s(p) = \min\{\max\{p, \tau_{min}\}, \tau_{\max}\} \tag{6}$$

In what follows, we consider steady problems of (slow) Bingham flow with pressure dependent yield stress that satisfies

$$\begin{cases} -\nabla \cdot \boldsymbol{\tau} + \nabla p = 0 & \text{in } \Omega \\ \nabla \cdot \boldsymbol{u} = 0 & \text{in } \Omega \\ \boldsymbol{u} = \boldsymbol{g}_D & \text{on } \Gamma_D \end{cases} \tag{7}$$

and proceed within the framework of generalized Stokes problems. So, we introduce the second invariant of the strain rate tensor $\gamma_{\mathrm{II}} = \frac{1}{2}(2\mathbf{D} : 2\mathbf{D})$, resp., $\|\mathbf{D}\| = \frac{1}{\sqrt{2}}\gamma_{\mathrm{II}}^{\frac{1}{2}}$, and define a generalized viscosity $\eta(\cdot,\cdot)$ which depends on the pressure and the shear rate:

$$\eta(\gamma_{\mathrm{II}}, p) = \nu + \frac{\sqrt{2}}{2} \frac{\tau_s(p)}{\gamma_{\mathrm{II}}^{\frac{1}{2}}} \tag{8}$$

To define the viscosity everywhere, we introduce the classical regularization:

$$\eta(\gamma_{\mathrm{II}}, p) = \nu + \frac{\sqrt{2}}{2} \frac{\tau_s(p)}{(\gamma_{\mathrm{II}} + \epsilon^2)^{\frac{1}{2}}} \tag{9}$$

As a consequence, Bingham flow with pressure dependent yield stress is the limit case, $\epsilon = 0$, of the regularized problem. However, it is well known that the accuracy of the solution is strongly dependent on this parameter ϵ. Summarizing the previous considerations, the considered system of equations in the primitive variables $\boldsymbol{u}$ and p is given as follows:

$$\begin{cases} -\nabla \cdot (2\eta(\gamma_{\mathrm{II}}, p)\mathbf{D}(\boldsymbol{u})) + \nabla p = 0 & \text{in } \Omega \\ \nabla \cdot \boldsymbol{u} = 0 & \text{in } \Omega \\ \boldsymbol{u} = \boldsymbol{g}_D & \text{on } \Gamma_D \end{cases} \tag{10}$$

2 Non-standard Saddle Point Problem Formulation

After discretization, for instance with standard Q2P1 finite elements, let $\tilde{\boldsymbol{u}} = (\boldsymbol{u}, p)$ and $\mathscr{R}_{\tilde{\boldsymbol{u}}}$ denote the discrete residuals for the system (10). We use the Newton method which means that the nonlinear iteration is updated with the correction $\delta\tilde{\boldsymbol{u}}$, $\tilde{\boldsymbol{u}}^{n+1} = \tilde{\boldsymbol{u}}^n + \delta\tilde{\boldsymbol{u}}$. Then, the Newton linearization provides the following approximation for the residuals:

$$\begin{aligned} \mathscr{R}(\tilde{\boldsymbol{u}}^{n+1}) =& \mathscr{R}(\tilde{\boldsymbol{u}}^n + \delta\tilde{\boldsymbol{u}}) \\ \simeq& \mathscr{R}(\tilde{\boldsymbol{u}}^n) + \left[\frac{\partial \mathscr{R}(\tilde{\boldsymbol{u}}^n)}{\partial \tilde{\boldsymbol{u}}}\right] \delta\tilde{\boldsymbol{u}} \end{aligned} \tag{11}$$

Hence, one iteration of the Newton method can be written as follows:

$$\begin{bmatrix} \boldsymbol{u}^{n+1} \\ \\ p^{n+1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{u}^{n} \\ \\ p^{n} \end{bmatrix} - \omega_n \begin{bmatrix} \dfrac{\partial \mathscr{R}_{\boldsymbol{u}}(\boldsymbol{u}^n, p^n)}{\partial \boldsymbol{u}} & \dfrac{\partial \mathscr{R}_{\boldsymbol{u}}(\boldsymbol{u}^n, p^n)}{\partial p} \\ \dfrac{\partial \mathscr{R}_{p}(\boldsymbol{u}^n, p^n)}{\partial \boldsymbol{u}} & \dfrac{\partial \mathscr{R}_{p}(\boldsymbol{u}^n, p^n)}{\partial p} \end{bmatrix}^{-1} \begin{bmatrix} \mathscr{R}_{\boldsymbol{u}}(\boldsymbol{u}^n, p^n) \\ \\ \mathscr{R}_{p}(\boldsymbol{u}^n, p^n) \end{bmatrix} \tag{12}$$

The damping parameter $\omega_n \in (0, 1]$ is typically chosen such that:

$$\begin{bmatrix} \mathscr{R}_{\boldsymbol{u}}(\boldsymbol{u}^{n+1}, p^{n+1}) \\ \mathscr{R}_{p}(\boldsymbol{u}^{n+1}, p^{n+1}) \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \boldsymbol{u}^{n+1} \\ p^{n+1} \end{bmatrix} \leq \begin{bmatrix} \mathscr{R}_{\boldsymbol{u}}(\boldsymbol{u}^{n}, p^{n}) \\ \mathscr{R}_{p}(\boldsymbol{u}^{n}, p^{n}) \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \boldsymbol{u}^{n} \\ p^{n} \end{bmatrix} \tag{13}$$

As we will demonstrate for the considered yield stress fluids, this damping parameter is not enough to ensure robust convergence. In what follows, we derive explicitly the Jacobian in order to segregate it into "bad" and "good" terms to get a robust nonlinear solver. The block matrices of the Jacobian are given as follows:

$$\begin{aligned} \left[\frac{\partial \mathscr{R}_{\boldsymbol{u}}(\boldsymbol{u}^n, p^n)}{\partial \boldsymbol{u}}\right] \boldsymbol{v} = &- \nabla \cdot \Big(2\eta(\gamma_{\mathrm{II}}^n, p^n)\mathbf{D}(\boldsymbol{v}) \\ &+ 8\eta_1'(\gamma_{\mathrm{II}}^n, p^n)\left[\mathbf{D}(\boldsymbol{u}^n) : \mathbf{D}(\boldsymbol{v})\right] \mathbf{D}(\boldsymbol{u}^n)\Big) \end{aligned} \tag{14}$$

where $\eta_1'(\gamma_{\mathrm{II}}, p) = \frac{\partial \eta(\gamma_{\mathrm{II}}, p)}{\partial \gamma_{\mathrm{II}}}$, the last term in the equation (14), is due to the shear dependent viscosity models. Furthermore, there holds

$$\left[\frac{\partial \mathscr{R}_{\boldsymbol{u}}(\boldsymbol{u}^n, p^n)}{\partial p}\right] q = \Big(\mathbf{I} - 2\eta_2'(\gamma_{\mathrm{II}}^n, p^n)\mathbf{D}(\boldsymbol{u}^n)\Big) \nabla q \tag{15}$$

where $\eta_2'(\gamma_{\mathrm{II}}, p) = \frac{\partial \eta(\gamma_{\mathrm{II}}, p)}{\partial p}$, the second term in the equation (15), is relevant for pressure-dependent viscosity models. Moreover, the incompressibility condition leads to

$$\left[\frac{\partial \mathscr{R}_{p}(\boldsymbol{u}^n, p^n)}{\partial \boldsymbol{u}}\right] \boldsymbol{v} = -\nabla \cdot \boldsymbol{v} \tag{16}$$

and additionally we obtain:

$$\left[\frac{\partial \mathscr{R}_{p}(\boldsymbol{u}^n, p^n)}{\partial p}\right] q = 0 \tag{17}$$

Let $\mathbf{V}$ and $\mathbf{Q}$ be the spaces for velocity and pressure respectively, and let $\mathbf{V}'$ and $\mathbf{Q}'$ be their cooresponding dual spaces. The weak formulation reads:

$$\int_\Omega \left[\frac{\partial \mathscr{R}_{\boldsymbol{u}}(\boldsymbol{u}^n, p^n)}{\partial \boldsymbol{u}}\right] \boldsymbol{u} \cdot \boldsymbol{v} dx = \int_\Omega 2\eta(\gamma^n_{\mathrm{II}}, p^n)\,[\mathbf{D}(\boldsymbol{u}) : \mathbf{D}(\boldsymbol{v})]\,dx \\ + \int_\Omega 8\eta_1'(\gamma^n_{\mathrm{II}}, p^n)\,[\mathbf{D}(\boldsymbol{u}^n) \otimes \mathbf{D}(\boldsymbol{u})] : [\mathbf{D}(\boldsymbol{u}^n) \otimes \mathbf{D}(\boldsymbol{v})]\,dx \tag{18}$$

Next, let us introduce the following linear forms defined on $\mathbf{V} \longrightarrow \mathbf{V}'$

$$\begin{aligned} \langle \mathbf{A}_1 \boldsymbol{u}, \boldsymbol{v}\rangle &:= \int_\Omega 2\eta(\gamma^n_{\mathrm{II}}, p^n)\,[\mathbf{D}(\boldsymbol{u}) : \mathbf{D}(\boldsymbol{v})]\,dx \\ \langle \mathbf{A}_2 \boldsymbol{u}, \boldsymbol{v}\rangle &:= \int_\Omega 8\eta_1'(\gamma^n_{\mathrm{II}}, p^n)\,[\mathbf{D}(\boldsymbol{u}^n) \otimes \mathbf{D}(\boldsymbol{u})] : [\mathbf{D}(\boldsymbol{u}^n) \otimes \mathbf{D}(\boldsymbol{v})]\,dx \\ \langle \mathbf{A} \boldsymbol{u}, \boldsymbol{v}\rangle &:= \langle \mathbf{A}_1 \boldsymbol{u}, \boldsymbol{v}\rangle + \langle \mathbf{A}_2 \boldsymbol{u}, \boldsymbol{v}\rangle \end{aligned} \tag{19}$$

and the associated bilinear forms defined on $\mathbf{V} \times \mathbf{V} \longrightarrow \mathbb{R}$

$$a(\boldsymbol{u}, \boldsymbol{v}) = \langle \mathbf{A}\boldsymbol{u}, \boldsymbol{v}\rangle, \quad a_1(\boldsymbol{u}, \boldsymbol{v}) = \langle \mathbf{A}_1\boldsymbol{u}, \boldsymbol{v}\rangle, \quad a_2(\boldsymbol{u}, \boldsymbol{v}) = \langle \mathbf{A}_2\boldsymbol{u}, \boldsymbol{v}\rangle \tag{20}$$

and the linear forms defined on $\mathbf{V} \longrightarrow \mathbf{Q}'$:

$$\langle \mathbf{B}\boldsymbol{u}, p\rangle := -\int_\Omega \nabla\cdot\boldsymbol{u}\ p\,dx \tag{21}$$

the new additional linear forms $\tilde{\mathbf{B}}$ and $\mathbf{C}$ are given as follows

$$\langle \tilde{\mathbf{B}}\boldsymbol{u}, p\rangle = \int_\Omega \nabla \cdot \Big(2\eta_2'(\gamma^n_{\mathrm{II}}, p^n)\mathbf{D}(\boldsymbol{u}^n)\boldsymbol{u}\Big)\ p\,dx \tag{22}$$

$$\langle \mathbf{C}\boldsymbol{u}, p\rangle = -\int_\Omega \nabla \cdot \Big[\Big(\mathbf{I} - 2\eta_2'(\gamma^n_{\mathrm{II}}, p^n)\mathbf{D}(\boldsymbol{u}^n)\Big)\boldsymbol{u}\Big]\ p\,dx \tag{23}$$

with the associated bilinear forms $b(\cdot,\cdot)$, $\tilde{b}(\cdot,\cdot)$, and $c(\cdot,\cdot)$ defined on $\mathbf{V} \times \mathbf{Q} \longrightarrow \mathbb{R}$ read:

$$b(\boldsymbol{v}, q) = \langle \mathbf{B}\boldsymbol{v}, q\rangle\,, \quad \tilde{b}(\boldsymbol{v}, q) = \langle \tilde{\mathbf{B}}\boldsymbol{v}, q\rangle, \quad c(\boldsymbol{v}, q) = b(\boldsymbol{v}, q) + \tilde{b}(\boldsymbol{v}, q) \tag{24}$$

So, the corresponding Newton iteration (12) after discretization becomes:

$$\begin{bmatrix} \boldsymbol{u}^{n+1} \\ p^{n+1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{u}^{n} \\ p^{n} \end{bmatrix} - \omega_n \begin{bmatrix} \mathbf{A} & \mathbf{C}^{\mathrm{T}} \\ \mathbf{B} & 0 \end{bmatrix}^{-1} \begin{bmatrix} \mathscr{R}_{\boldsymbol{u}}(\boldsymbol{u}^n, p^n) \\ \mathscr{R}_{p}(\boldsymbol{u}^n, p^n) \end{bmatrix} \tag{25}$$

In the case of pressure-dependent yields stress, the Jacobian has a nonsymmetric saddle point structure (if not, then $\mathbf{C}^{\mathrm{T}} = \mathbf{B}^{\mathrm{T}}$):

$$\mathbf{J} = \begin{bmatrix} \mathbf{A} & \mathbf{C}^{\mathrm{T}} \\ \mathbf{B} & 0 \end{bmatrix} \tag{26}$$

The Jacobian $\mathbf{J}$ can be decomposed, based on the block operators $\mathbf{A}$, into

$$\mathbf{A} = \mathbf{A}_1 + \mathbf{A}_2 \tag{27}$$

respectively, $\mathbf{C}$ into

$$\mathbf{C} = \mathbf{B} + \tilde{\mathbf{B}}. \tag{28}$$

In what follows, we mainly concentrate, first of all, onto the model (1) for yield stress only and take $\mathbf{C} = \mathbf{B}$ (at the end, we provide a preliminary result for pressure-dependent viscosity, too). Therefore, our studies focus on the discussed decomposition of the operator $\mathbf{A}$ due to (27).

3 Robust Nonlinear Solver

To develop a robust nonlinear solver we introduce a new control parameter δ_n in order to balance the operators $\mathbf{A}_1$ (corresponding to the typical fixed point approach) and $\mathbf{A}_2$, both being part of the complete Jacobian $\mathbf{A}$:

$$\mathbf{A} = \mathbf{A}_1 + \delta_n \mathbf{A}_2 \tag{29}$$

In the present note, we concentrate on the choice of the optimal parameter δ_n balancing the fixed point and the full Newton iteration. We take the classical flow around cylinder benchmark [5], and perform corresponding simulations for Bingham flow.

First, we take a very small yield stress parameter, $\tau_s = 10^{-4}$, and apply the fixed point ($\delta_n = 0$) and classical Newton ($\delta_n = 1$) methods. Table 1 shows the resulting numbers of nonlinear iterations. Both methods, Newton and fixed point, are easily converging towards the solution, more or less independent of the mesh level. Moreover, the Newton method overcomes the fixed point method, as expected, due to the moderate nonlinearity. To highlight the insufficiency of the globally damped Newton (13) to simulate Bingham flow problems, we further increase the yield stress. Now, the Newton method can only converge with a strong damping parameter ω_n as the yield stress increases, for instance $\omega_n = 0.1$ for $\tau_s = 10^{-2}$, and no convergence at all can be obtained for higher yield stress, $\tau_s \geq 10^{-1}$. Instead,

Table 1 Globally damped Newton: The numbers of nonlinear iterations for Bingham flow with fixed point method and globally damped Newton for increasing yield stress

	$\tau_s = 10^{-4}$		$\tau_s = 10^{-3}$		$\tau_s = 10^{-2}$		$\tau_s = 10^{-1}$	$\tau_s = 1$
	Fixed	Newton	Fixed	Newton	Fixed	Newton	Fixed	Fixed
Level	point	$\omega_n = 1.0$	point	$\omega_n = 0.2$	point	$\omega_n = 0.1$	point	point
2	21	3	67	99	212	210	490	1032
3	24	5	84	95	308	200	728	2135
4	20	5	98	90	408	190	1375	3444

Table 2 Statically balanced Newton: The numbers of nonlinear iterations for Bingham flow with two different yield stress values $\tau_s = 10^{-2}$ and $\tau_s = 10^{-1}$, with a statically balanced Jacobian, i.e. δ_n is kept constant

τ_s	Level	$\delta_n = 0.1$	$\delta_n = 0.25$	$\delta_n = 0.5$	$\delta_n = 0.6$
10^{-2}	2	236	198	135	110
	3	352	295	199	160
	4	455	380	256	206
10^{-1}	2	551	461	311	251
	3	848	708	475	382
	4	1455	1214	813	653

the fixed point method can converge for all cases, however being very slow and not being robust w.r.t. mesh level and/or yield stress.

Clearly, with increasing yield stress, it is hard if not impossible to solve the corresponding flow problems with the globally damped Newton. Therefore, in the next step, we take a static δ_n, i.e. $\delta_n = \delta_0$ for $n \geq 1$, which has been introduced in (29). The balancing parameter δ_n is taken as a constant increasing from 0 to 1. Table 2 presents the numbers of nonlinear iterations for Bingham flow with different values for the yield stress. From the results in Table 2, it is clear that increasing the contribution from the operator $\mathbf{A}_2$ improves the convergence behavior, but this contribution needs to remain under control. To do so, we go for a dynamic change of δ_n w.r.t. the residual changes. From the numerical experiment it can be noticed that the dynamic changes of the residual give a precious information about the singularity of the Jacobian. Indeed, the larger relative changes in the residual with the operator $\mathbf{A}_1$ reflect the 'singularity' of the operator $\mathbf{A}_2$. In this case, the parameter δ_n should have a small relative change and remain small. Moreover, when the relative changes in the residual are close to zero, this indicates that the operator $\mathbf{A}_2$ has the nicest properties and δ_n can be increased accordingly and maintained close to 1. We introduce the increment

$$\mathscr{Q}_n := \frac{\left\|\mathscr{R}(\gamma_{\text{II}}^n, p^n)\right\|}{\left\|\mathscr{R}(\gamma_{\text{II}}^{n-1}, p^{n-1})\right\|}, \tag{30}$$

Table 3 **Behavior of the weighted Newton w.r.t. starting parameter**: The numbers of nonlinear iterations for the dynamically balanced Newton for Bingham flow for a wide range of yield stress values, varying from 10^{-3} to 5, for different initial values δ_0

	τ_s						
δ_0	0.001	0.01	0.1	0.5	1.0	2.0	5.0
0.0	10	15	20	19	19	20	20
0.3	10	16	20	19	19	20	20
0.7	18	18	22	22	20	18	18
1.0	46	14	19	21	21	22	22

Table 4 **Convergence w.r.t. continuation strategies**: The numbers of nonlinear iterations for the dynamically balanced Newton for Bingham flow for increasing yield stress values, from 10^{-3} to 5, and decreasing ϵ, from 10^{-2} to 10^{-5}

	τ_s						
ϵ	0.001	0.01	0.1	0.5	1.0	2.0	5.0
Continuation Newton w.r.t. ϵ							
10^{-2}	10	15	20	19	19	20	20
10^{-3}	11	11	12	17	16	15	15
10^{-4}	15	13	18	16	15	26	15
10^{-5}	16	10	22	22	17	15	17
Continuation Newton w.r.t. τ_s							
10^{-2}	10	14	19	12	8	7	7
10^{-3}	14	20	26	15	8	8	8
10^{-4}	21	26	34	23	10	17	8
10^{-5}	22	45	41	29	11	10	10

and define the following continuous function for changes of δ_n w.r.t. the residual $\mathscr{R}_n$:

$$\frac{\delta_{n+1}}{\delta_n} = 0.2 + \frac{4}{0.7 + \exp(1.5\mathscr{Q}_n)} \tag{31}$$

It should be pointed out that the choice (31) of δ_n is derived so far based on simple and preliminary numerical experiments only. We check the robustness of the dynamic changes of δ_n in (31) for various values of yield stress. Table 3 shows the numbers of nonlinear iterations for Bingham flow for a wide range of yield stress values and different starting weighting factors for the Jacobian, that means δ_0.

Since the convergence typically gets harder with smaller values for the regularization parameter ϵ, we check the robustness of the dynamic changes of δ_n in (31) for decreasing ϵ and a wide range of yield stress values. Table 4 shows the numbers of nonlinear iterations for Bingham flow using continuation strategies w.r.t. ϵ as well as w.r.t. τ_s.

Table 5 **Pressure dependent yield stress**: The numbers of nonlinear iterations for Bingham flow with pressure dependent yield stress in (6) with fixed point method $\delta_n = 0.0$ and dynamically balanced Newton, varying the lower bound yield stress τ_{min} and fixed upper bound yield stress $\tau_{max} = 0.1$

	τ_{min}				
Method	0.0	0.0001	0.001	0.01	0.1
Fixed point	356	356	356	356	356
Newton	79	68	68	56	27

Moreover, it should be pointed out that the parameters ϵ and τ_s can be seen as bounds for some physical quantities in models for granular material [2, 3].

Finally, we want to perform some preliminary tests regarding the flexibility and robustness of the dynamically balanced Newton method for pressure-dependent yield stress. In a first step, the yield stress is taken as a function of the pressure as described in (6). We fix the upper bound of the yield stress and change the lower bound of the yield stress to allow significant changes in the pressure which should mainly influence the convergence behavior. However, due to the 'min-max' flow rule, we cannot differentiate w.r.t. the pressure so that we apply the described Newton modification for the velocity part only, while the pressure dependence is treated in a fixed point style only. Nevertheless, the comparison of the standard fixed point method and the newly dynamically balanced Newton, which is presented in Table 5, shows already a clearly improved behavior. In the next step, we will apply a flow model including pressure and shear rate which will allow differentiation w.r.t. both arguments (as demonstrated in the shown models for granular flow) so that an extension of the new Newton method to pressure-dependent yield stress fluids can be realized and numerically analyzed, too.

4 Summary

We shortly presented a new Newton-type method for flow problems with yield stress which are typical for viscoplastic Bingham models as well as granular flow models with pressure-dependent yield stress. The model is approximated with a regular approach to derive the Jacobian. Then, the partial contributions to the Jacobian are segregated in order to differ between 'good' and 'bad' parts (due to their expected numerical behavior). Firstly, we showed the insufficiency of the classical globally damped Newton. Secondly, we derived a statically balanced Newton approach, by taking different parts of the Jacobian in a static manner for different yield stress values. Thirdly, we went further with dynamic changes allowing the selection of the 'optimal' contributions inside of the Jacobian, here mainly based on the residual changes. The numerical results demonstrate the ability to simulate the Bingham viscoplastic model in the primitive variables for a small regularized parameter ϵ and pressure-dependent yield stress. Moreover, we pointed out how this approach can be extended to more complex (and more realistic) flow models which are typical for granular flow models.

Acknowledgements This work was supported by DAAD and the grant STW/DFG:TU 102/44-1.

References

1. N. El Khouja, N. Roquet, B. Cazacliu, Analysis of a regularized bingham model with pressure-dependent yield stress. J. Math. Fluid Mech. **17**, 723–739 (2015)
2. P. Jop, Y. Forterre, O. Pouliquen, A constitutive law for dense granular flows. Nature **441**, 727–730/28–55 (2006)
3. M.K. Langroudi, S. Turek, A. Ouazzi, G. Tardos, An investigation of frictional and collisional powder flows using a unified constitutive equation. Powder Technol. **197**, 91–101 (2009)
4. D.G. Schaeffer, Instability in the evolution equation describing incompressible granular flow. J. Differ. Equ. **66**, 19–50/28–55 (1987)
5. S. Turek, M. Schäfer, Efficient solvers for incompressible flow problems: an algorithmic and computational approach. Notes Num. Fluid Mech. **52**, 547–566/28–55 (1996)

Numerical Simulation of 3D Flow of Viscous and Viscoelastic Fluids in T-Junction Channel

Radka Keslerová and David Trdlička

Abstract This paper is interested in the numerical simulation of steady flows of laminar incompressible viscous and viscoelastic fluids through the channel with T-junction. The flow is described by the system of generalized incompressible Navier-Stokes equations. For the different choice of fluids model the different model of the stress tensor is used, Newtonian and Oldroyd-B models. Numerical tests are performed on three dimensional geometry, a branched channel with one entrance and two outlet parts. Numerical solution of the described models is based on cell-centered finite volume method using explicit Runge-Kutta time integration.

1 Introduction

Branching of pipes occurs in many technical or biological applications. In [13] the effects of viscoelasticity on the pitchfork bifurcation using a numerical finite volume method was investigated. Results from both the upper-convected Maxwell and Oldroyd-B models show that the instability occurs at lower Reynolds numbers for viscoelastic fluids in comparison to the Newtonian base case. In [12] computational fluid dynamics simulations of steady viscoelastic flows through a planar two dimensional T-junction is considered and the influence of constitutive model and fluid elasticity upon the main recirculating flow characteristics formed at the junction and the shear stress fields is studied. In biomedical applications, it is the complex branching system of blood vessels in human body. The blood can be characterized by shear-thinning viscoelastic property and the blood flow can be described by generalized Oldroyd-B model. In [1] the new model to describe the rheological characteristics of blood (namely shear-thinning and deformation dependent viscoelasticity) in both steady and unsteady flows was developed. In [5] a comparative numerical study of non-Newtonian fluid models capturing shear-thinning and viscoelastic effects of blood flow in idealized and realistic stenosed vessels was presented.

R. Keslerová (✉) • D. Trdlička
Czech Technical University in Prague, Karlovo nám. 13, 121 35 Prague, Czech Republic
e-mail: radka.keslerova@fs.cvut.cz; david.trdlicka@fs.cvut.cz

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_47

In prewious work we studied the numerical simulation of generalized Newtonian and Oldroyd-B fluids flow in 2D branching channel, [7, 8]. In this article this problem will be extended to the study of generalized Newtonian and Oldroyd-B fluids flow in 3D branching channel with T-junction.

2 Mathematical Model

The governing system of equations is the system of generalized Navier-Stokes equations, see [2]. This system consists of the continuity equation

$$\operatorname{div} \boldsymbol{u} = 0 \tag{1}$$

and the momentum equations

$$\rho \frac{\partial \boldsymbol{u}}{\partial t} + \rho(\boldsymbol{u}.\nabla)\boldsymbol{u} = -\nabla P + \operatorname{div} \mathbf{T}, \tag{2}$$

where P is the pressure, ρ is the constant density, $\boldsymbol{u}$ is the velocity vector and by the symbol $\mathbf{T}$ the stress tensor is denoted.

In order to solve (2) a constitutive relation is needed for the stress tensor.

For Newtonian fluids the stress tensor $\mathbf{T}$ is modelled by (see e.g. [3, 5])

$$\mathbf{T} = 2\mu\mathbf{D}, \tag{3}$$

where μ is the dynamic fluid viscosity and tensor $\mathbf{D}$ is the symmetric part of the velocity gradient, $\mathbf{D} = \frac{1}{2}(\nabla\boldsymbol{u} + \nabla\boldsymbol{u}^T)$.

In the case of viscoelastic fluids, the simplest viscoelastic model (Maxwell model) reads

$$\mathbf{T} + \lambda_1 \frac{\delta \mathbf{T}}{\delta t} = 2\mu\mathbf{D}, \tag{4}$$

where λ_1 is the relaxation time [5]. The symbol $\frac{\delta}{\delta t}$ represents the upper convected derivative defined by the relation, for more details see [4, 5]

$$\frac{\delta \mathbf{T}_e}{\delta t} = \frac{\partial \mathbf{T}_e}{\partial t} + (\boldsymbol{u}.\nabla)\mathbf{T}_e - (\mathbf{W}\mathbf{T}_e - \mathbf{T}_e\mathbf{W}) - (\mathbf{D}\mathbf{T}_e + \mathbf{T}_e\mathbf{D}), \tag{5}$$

where $\mathbf{D}$ is symmetric and $\mathbf{W}$ is antisymmetric part of the velocity gradient, $\mathbf{W} = \frac{1}{2}(\nabla\boldsymbol{u} - \nabla\boldsymbol{u}^T)$.

By combination of these two mathematical models (3) and (4) (Newtonian and Maxwell) the behaviour of mixture of viscous and viscoelastic fluids can be described. This model is called Oldroyd-B and it has the form

$$\mathbf{T} + \lambda_1 \frac{\delta \mathbf{T}}{\delta t} = 2\mu \left(\mathbf{D} + \lambda_2 \frac{\delta \mathbf{D}}{\delta t} \right), \tag{6}$$

where symbols λ_1 is relaxation time and λ_2 is the retardation time (with dimension of time).

For the numerical modelling of the generalized Newtonian and Oldroyd-B fluids flow it is necessary to generalize the mathematical models. The constant viscosity coefficient μ is replaced by a shear rate dependent viscosity function $\mu(\dot{\gamma})$ where shear rate $\dot{\gamma}$ is defined by $\dot{\gamma} = 2\sqrt{\frac{1}{2}\mathrm{tr}\,\mathbf{D}^2}$. This function can be written in the following general form (for more details see [14, 15])

$$\mu(\dot{\gamma}) = \mu_\infty + \frac{\mu_0 - \mu_\infty}{(1 + (\lambda\dot{\gamma})^b)^a}, \tag{7}$$

the following parameters found in [9] have been used for the flow simulations presented in this paper:

$\mu_0 = 1.6 \cdot 10^{-1}$ Pa s, $\mu_\infty = 3.6 \cdot 10^{-3}$ Pa s, $a = 1.23, b = 0.64, \lambda = 8.2$ s.

To account for the viscoelasticity of fluids flow the equations for the conservation of linear momentum 1 and mass 2 is considered where the extra stress tensor $\mathbf{T}$ is decomposed into its Newtonian part $\mathbf{T}_s$ and its elastic part $\mathbf{T}_e$, $\mathbf{T} = \mathbf{T}_s + \mathbf{T}_e$, such that

$$\mathbf{T}_s = 2\mu_s \mathbf{D} \tag{8}$$

and $\mathbf{T}_e$ satisfies a constitutive equation of Oldroyd-B type, namely

$$\frac{\partial \mathbf{T}_e}{\partial t} + (\boldsymbol{u}.\nabla)\mathbf{T}_e = \frac{2\mu_e}{\lambda_1}\mathbf{D} - \frac{1}{\lambda_1}\mathbf{T}_e + (\mathbf{W}\mathbf{T}_e - \mathbf{T}_e\mathbf{W}) + (\mathbf{D}\mathbf{T}_e + \mathbf{T}_e\mathbf{D}), \tag{9}$$

the μ_e denotes the elastic viscosity coefficient and λ_1 is the relaxation time, $\mu_e = 4.0 \cdot 10^{-1}$ Pa s and $\lambda_1 = 0.06$ s (according to [9]). The following four special parameters settings related to four specific models will be used in our numerical simulations:

Newtonian	$\mu_s(\dot{\gamma}) = \mu_\infty$	$\mathbf{T}_e \equiv 0$
Generalized Newtonian	$\mu_s(\dot{\gamma})$	$\mathbf{T}_e \equiv 0$
Oldroyd-B	$\mu_s(\dot{\gamma}) = \mu_\infty$	$\mathbf{T}_e$
Generalized Oldroyd-B	$\mu_s(\dot{\gamma})$	$\mathbf{T}_e$

3 Numerical Solution

The mathematical models described above are solved numericaly the artificial compressibility approach combined with the finite-volume discretization. The artificial compressibility method [6, 8, 11] is used to obtain equation for pressure. It means that the continuity equation is completed by a pressure time derivative term $\frac{\partial p}{\beta^2 \partial t}$, where β is positive parameter, making the inviscid part of the system of equations hyperbolic

$$\frac{1}{\beta^2}\frac{\partial p}{\partial t} + \operatorname{div} \boldsymbol{u} = 0, \tag{10}$$

the parameter β in this work is chosen equal to the maximum inlet velocity. This value ensures good convergence to steady state but is not large enough to make the transient solution accurate in time. Therefore it is suitable for steady flows only. The discretization is done by a cell-centered finite-volume method with hexahedral finite volumes. The system including the modified continuity equation and the momentum equations can be written

$$\tilde{R}_\beta W_t + F_x^c + G_y^c + H_z^c = F_x^v + G_y^v + H_z^v + S, \quad \tilde{R}_\beta = \operatorname{diag}(\frac{1}{\beta^2}, 1, \cdots, 1), \tag{11}$$

where W is vector of unknowns, $W = (p, u, v, w, t_{e1}, \ldots, t_{e6})$, by superscripts c and v the inviscid and the viscous fluxes are denoted. The symbol S denotes the source term.

Equation (11) is discretized in space by the finite volume method and the arising system of ODEs is integrated in time by the explicit multistage Runge–Kutta scheme [8, 10].

The flow is modelled in a bounded computational domain where a boundary is divided into three mutually disjoint parts: a solid wall, an outlet and an inlet. At the inlet Dirichlet boundary condition for velocity vector and for the stress tensor is used and for the pressure homogeneous Neumann boundary condition is used. At the outlet parts the pressure value is prescribed and for the velocity vector and the stress tensor homogeneous Neumann boundary condition is used. The no-slip boundary condition for the velocity vector is used on the wall. For the pressure and stress tensor homogeneous Neumann boundary condition is considered.

4 Numerical Results

This section deals with the comparison of the numerical results of generalized Newtonian and generalized Oldroyd-B fluids flow. Numerical tests are performed in an idealized branched channel with the circle cross-section. Figure 1 (left) shows

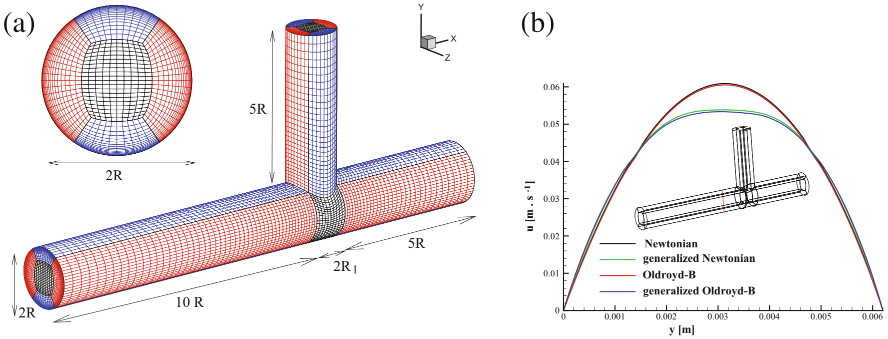


Fig. 1 Structure of the tested domain (*left*) and axial velocity profile of tested fluids (*right*). (**a**) Structure of the domain. (**b**) Axial velocity profile

the shape of the tested domain. The computational domain is discretized using a structured, wall fitted mesh with hexahedral cells. The domain is divided to 19 blocks with 125,000 cells.

All numerical results presented in this section were computed using in-house software. The computational code was verified for the steady flow of an incompressible fluid in a straight tube by prescribing a constant velocity profile at the inlet. Fully developed velocity profiles for all tested fluids were used as the initial velocity profile for following computations in the branching channel with T-junction. At the outlet the constant pressure values are prescribed (0.0005 Pa (main channel) and 0.00025 Pa (branch)).

The computations were performed with the following model parameters: $R = 0.0031\,\mathrm{m}$, $R_1 = 0.0025\,\mathrm{m}$, $\mu_e = 0.0004\,\mathrm{Pa\ s}$, $\mu_s = 0.0036\,\mathrm{Pa\,s}$, $\lambda_1 = 0.06\,\mathrm{s}$, $U_0 = 0.0615\,\mathrm{m\,s^{-1}}$, $\rho = 1050\,\mathrm{kg\,m^{-3}}$. In Fig. 1 the axial velocity profile for tested types of fluids close to the branching is shown. The lines for Newtonian and Oldroyd-B fluids are similar to the parabolic line, as was assumed. From this velocity profile is clear that the shear thinning fluids attain lower maximum velocity in the central part of the channel (close to the axis of symmetry) which is compensated by the increase of local velocity in the boundary layer close to the wall.

In Fig. 2 the velocity isolines and the cuts through the main channel and the small branch are shown.

The axial velocity isolines in the center-plane area for all tested fluids are shown in the Fig. 3. It can be observed from these that the size of separation region for generalized Newtonian and generalized Oldroyd-B fluids is smaller than for Newtonian and Oldroyd-B fluids, see in detail Fig. 4.

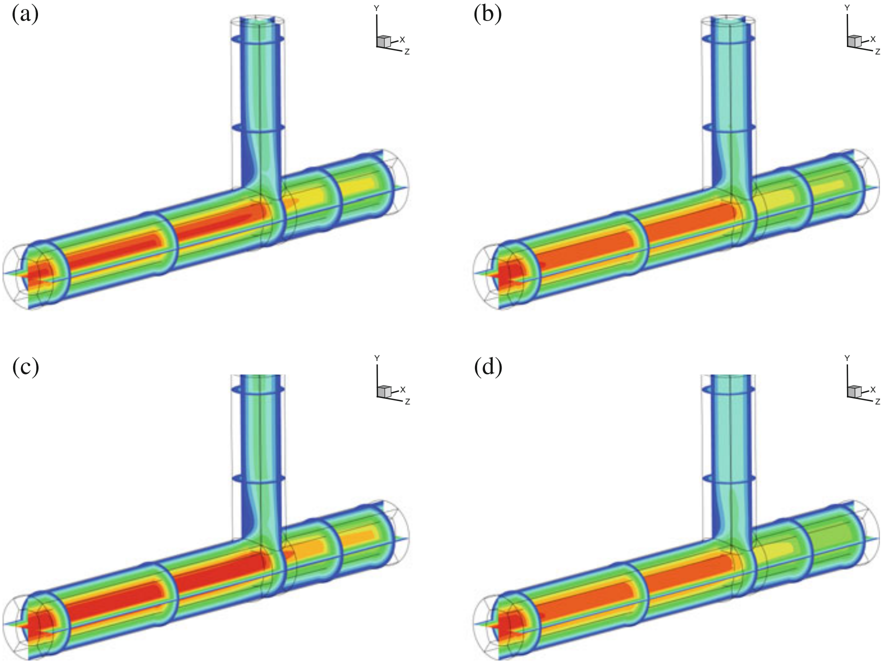


Fig. 2 Velocity isolines of steady flows for generalized Newtonian and Oldroyd-B fluids. (**a**) Newtonian. (**b**) Generalized Newtonian. (**c**) Oldroyd-B. (**d**) Generalized Oldroyd-B

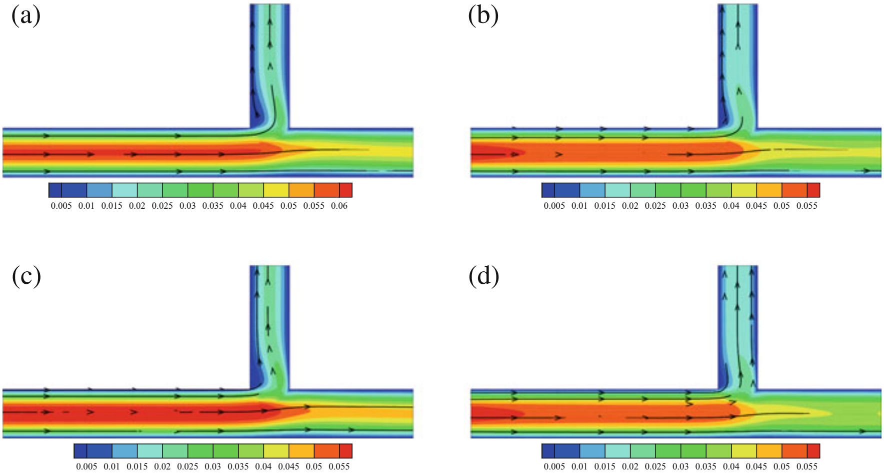


Fig. 3 Axial velocity isolines in the center-plane area for generalized Newtonian and Oldroyd-B fluids. (**a**) Newtonian. (**b**) Generalized Newtonian. (**c**) Oldroyd-B. (**d**) Generalized Oldroyd-B

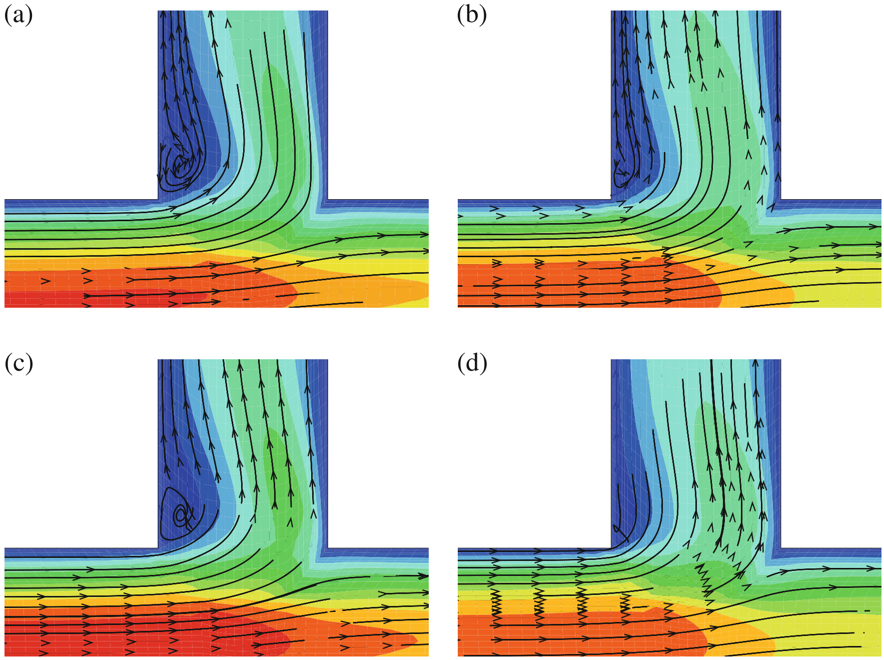


Fig. 4 Axial velocity isolines in the center-plane area in the separation region. (**a**) Newtonian. (**b**) Generalized Newtonian. (**c**) Oldroyd-B. (**d**) Generalized Oldroyd-B

In the table the ratios of volume flow rates in the branches are presented. These ratios are given by Q_1/Q_2 where Q_1 is the volume flow rate in the small branch and Q_2 is the volume flow rate in the main channel (at the output).

	Newtonian	gen. Newtonian	Oldroyd-B	gen. Oldroyd-B
Ratio	0.781498	0.821212	0.779099	0.852804

5 Conclusions

Classical Newtonian and Oldroyd-B models, as well as their generalized (shear-thinning) modifications have been considered to model flow in the branching channel with T-junction, to investigate shear-thinning and viscoelastic effects in steady flow simulations. Based on the above computation results we conclude that in this type of the channel the numerical results for Newtonian and Oldroyd-B fluids are similar. From the presented velocity profile is clear that the shear thinning fluids (generalized Newtonian and Oldroyd-B fluids) attain lower maximum velocity in the central part of the channel (close to the axis of symmetry) which is compensated by the increase of local velocity in the boundary layer close to the wall.

The numerical method used to solve the governing equations seems to be sufficiently robust and efficient for the appropriate resolution of the given class of problems.

The future work will be occupy with the numerical simulation of unsteady flows for viscous and viscoelastic fluids in the three dimensional branching channel with circle cross section.

Acknowledgements This work was supported by grant SGS16/206/OHK2/3T/12 of the Czech Science Foundation.

References

1. M. Anand, J. Kwack, A. Masud, A new generalized Oldroyd-B model for blood flow in complex geometries. Int. J. Eng. Sci. **72**, 78–88 (2013)
2. L. Beneš, P. Louda, K. Kozel, R. Keslerová, J. Štigler, *Numerical simulations of flow through channels with T-junction*. Appl. Math. Comput. **219**, 7225–7235 (2013)
3. T. Bodnar, A. Sequeira, Numerical study of the significance of the non-Newtonian nature of blood in steady flow through stenosed vessel, in *Advances in Mathematical Fluid Mechanics* (Springer, Heidelberg/Dordrecht, 2010), pp. 83–104
4. T. Bodnar, A. Sequeira, L. Pirkl, Numerical simulations of blood flow in a stenosed vessel under different flow rates using a generalized Oldroyd-B model. Numer. Anal. Appl. Math. **2**, 645–648 (2009)
5. T. Bodnar, A. Sequeira, M. Prosi, On the shear-thinning and viscoelastic effects of blood flow under various flow rates. Appl. Math. Comput. **217**, 5055–5067 (2010)
6. A.J. Chorin, A numerical method for solving incompressible viscous flow problem. J. Comput. Phys. **135**, 118–125 (1967)
7. R. Keslerová, K. Kozel, Numerical simulation of viscous and viscoelastic fluids flow by finite volume method, in *6th International Symposium on Finite Volumes for Complex Applications*, Prague (2011)
8. R. Keslerová, K. Kozel, Numerical simulation of generalized Newtonian and Oldroyd-B fluids flow, in *16th Seminar on Programs and Algorithms of Numerical Mathematics*, Dolní Maxov (2012)
9. A. Leuprecht, K. Perktold, Computer simulation of non-Newtonian effects of blood flow in large arteries. Comput. Methods Biomech. Biomech. Eng. **4**, 149–163 (2001)
10. R. LeVeque, *Finite-Volume Methods for Hyperbolic Problems* (Cambridge University Press, Cambridge/New York, 2004)
11. P. Louda, J. Příhoda, K. Kozel, P. Sváček, Numerical simulation of flows over 2D and 3D backward-facing inclined steps. Int. J. Heat Fluid Flow **43**, 268–276 (2013)
12. H.M. Matos, P.J. Oliveira, Steady flows of constant-viscosity viscoelastic fluids in a planar T-junction. J. Non-Newton. Fluid Mech. **213**, 15–26 (2014)
13. R.J. Poole, S.J. Haward, M.A. Alves, Symmetry-breaking bifurcations in T-channel flows: effects of fluid viscoelasticity. Procedia Eng. **79**, 28–34 (2014)
14. M.G. Rabby, A. Razzak, M.M. Molla, Pulsatile non-Newtonian blood flow through a model of arterial stenosis. Procedia Eng. **56**, 225–231 (2013)
15. J. Vimmr, A. Jonášová, O. Bublík, Effects of three geometrical parameters on pulsatile blood flow in complete idealised coronary bypasses. Math. Comput. Simul. **80**, 1324–1336 (2010)

Computational Simulations of Fractional Flow Reserve Variability

Timur Gamilov, Philippe Kopylov, and Sergey Simakov

Abstract Fractional flow reserve (FFR) is the golden standard for making decision on surgical treatment of coronary vessels with multiple stenosis. Clinical measurements of FFR require expensive invasive procedure with endovascular ultrasound probe. In this work a method of FFR simulation is considered. It is based on modelling 1D haemodynamics in patient-specific coronary vessels network reconstructed from CT scans. In contrast to our previous studies we used explicit minimum oscillating 2nd order characteristic method for internal nodes and 2nd order approximation of compatibility conditions for discretization of boundary conditions in junctions. The model is applied for simulating the change of FFR due to variability of the vessels elasticity and autoregulation response rate.

1 Introduction

Multiple stenosis of coronary arteries is a common cardiovascular disease. It can cause myocard ischemia, which frequently results in disability or death. Stenosis is usually treated invasively by stenting or noninvasively by drugs therapy. The choice is based on the estimate of haemodynamical importance of the stenosis. The modern criterion of the coronary stenosis severity is fractional flow reserve (FFR) [5]. FFR is calculated as a ratio of mean pressure distal to stenosis to mean aortic pressure during vasodilator administration [5]. The values below 0.8 associated with haemodynamic importance of the stenosis and recommendation of surgical treatment. The FFR based decisions allows reducing the number of unnecessary invasive treatment [10].

T. Gamilov (✉) • S. Simakov
Moscow Institute of Physics and Technology, Dolgoprudny, Russia

Institute of Numerical Mathematics RAS, Moscow, Russia
e-mail: gamilov@crec.mipt.ru

P. Kopylov
I.M. Sechenov First Moscow State Medical University, Moscow, Russia

Institute of Numerical Mathematics RAS, Moscow, Russia

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_48

Clinical measurements of FFR require expensive invasive procedure with endovascular ultrasound probe. Modern methods of FFR estimation involve 3D blood flow modelling in the local region of the studied vessel [5] (cFFR or virtual FFR). It requires complex simulations of local coronary region with high computational cost. Another approach is based on 1D haemodynamics simulation in coronary region [1, 2, 8]. It allows to simulate substantial part of coronary region with multiple stenoses. The 1D structure of coronary vessels can be reconstructed from patient's CT-scans [8]. In order to possibly decrease numerical resources needed for regular massive simulations in clinic we used in this work the explicit minimum oscillating second order characteristic method for internal nodes [4]. We present second order approximation of compatibility conditions along outgoing characteristics for discretization of boundary conditions in junctions. This numerical implementation is tested by grid refinement study (grid convergence).

The computational model is applied for simulating the change of FFR due to variability of the vessels elasticity and autoregulation response rate (ARR). These parameters may be affected by drugs administration, lifestyle or aging [6]. On the other hand, patient specific values cannot be estimated with sufficient accuracy. Thus, our method is a tool for an extended computational analysis of possible risk factors and stenosis severity for an individual patient. Such an analysis is impossible in general clinical studies.

2 Methods

The model of blood flow in coronary vascular network is based on unsteady viscous incompressible fluid flow through the 1D network of elastic tubes [7]. It takes into account patient-specific structure of coronary arteries. Systemic circulation is reduced to a single large vessel. Veins structure is supposed to be the same as for the arteries. The model is supplemented by autoregulation function according to [7]. In this section we briefly describe the model.

The flow in every vessel is described by the mass and momentum balance equations

$$\partial A_k/\partial t + \partial(A_k u_k)/\partial x = 0, \tag{1}$$

$$\partial u_k/\partial t + \partial\left(u_k^2/2 + p_k/\rho\right)/\partial x = f_{fr}(A_k, u_k), \tag{2}$$

where k is the index of the vessel; t is the time; x is the distance along the vessel counted from the vessel junction point; ρ is the blood density (constant); $A_k(t,x)$ is the vessel cross-section area; p_k is the blood pressure; $u_k(t,x)$ is the linear velocity averaged over the cross-section; f_{tr} is the friction force. Elasticity of the vessel wall is described by pressure-cross-section relation in the form

$$p_k(A_k) - p_{*k} = \rho c_k^2 f(A_k), \tag{3}$$

where $f(A)$ is monotone S-like function

$$f(A_k) = \begin{cases} \exp(\eta_k - 1) - 1, & \eta_k > 1 \\ \ln \eta_k, & \eta_k \leqslant 1, \end{cases} \tag{4}$$

p_{*k} is the pressure in the tissues surrounding the vessel; c_k is the velocity of small disturbances propagation in the wall; $\eta_k = A_k/A_{0k}$; A_{0k} is the unstressed cross-sectional area. At the vessels junctions the Poiseuille's pressure drop condition is applied

$$p_k(A_k(t,\tilde{x}_k)) - p^l_{node}(t) = \varepsilon_k R^l_k A_k(t,\tilde{x}_k)\, u_k(t,\tilde{x}_k), k = k_1, k_2, \ldots, k_M, \tag{5}$$

and combined with the compatibility conditions of hyperbolic set (1), (2) along outgoing characteristics and mass conservation condition.

At the terminal point of the venous system ($x = x_H$) the pressure $p_H = 8$ mmHg is set as the boundary condition. At the entry point of the aorta the blood flow is assigned as the boundary condition

$$u(t,0)\, S(t,0) = Q_H(t), \tag{6}$$

where $Q_H(t)$ is cardiac output profile for normal conditions (heart rate 1 Hz, stroke volume 65 ml [3]). Autoregulatory function is described as c_k adaptation for the changes of average pressure $\overline{p}_k$

$$\frac{c_{k,new}}{c_{k,old}} = \sqrt{\frac{\overline{p}_{k,new}}{\overline{p}_{k,old}}}, \tag{7}$$

where $\overline{p}_{k,new} = \dfrac{1}{(T_3 - T_2)l_k} \displaystyle\int_{T_2}^{T_3} \int_0^{l_k} p(x,t)dxdt$; l_k is the length of the k-th vessel; $\overline{p}_{k,old} = \dfrac{1}{(T_2 - T_1)l_k} \displaystyle\int_{T_1}^{T_2} \int_0^{l_k} p(x,t)dxdt$; T_1, T_2, T_3, T_4 are the initial moments of the successive averaging periods (successive cardiac cycles). The instant value of c_k is then calculated as

$$c_k = c_{k,old} + \gamma \frac{t - T_3}{T_4 - T_3}(c_{k,new} - c_{k,old}), \tag{8}$$

where $0 \leqslant \gamma \leqslant 1$ is the ARR. $\gamma = 1$ corresponds to the normal autoregulation, $\gamma = 0$ corresponds to the absence of autoregulation, $0 < \gamma < 1$ can be associated with vasodilator administration, etc.

An essential feature of coronary hemodynamics is partial compression of coronary arteries by myocard contraction during systole. It is simulated by setting $p_* = P^{cor}_{ext}(t)$ in (3). The shape of the function $P^{cor}_{ext}(t)$ is similar to the heart outflow profile [2]. Maximum value is set to 120 mmHg and 30 mmHg and applied for terminal vessels of the left and right coronary arteries. Increased resistance is simulated by multiplying R_k in (5) for all coronary vessels during systole by a factor of 3 [9].

2.1 Patient-Specific 1D Coronary Network

The computational domain in a 1D network was generated on the basis of patient-specific data. Processing of the CT scans requires complex algorithms described in [2]. Segmentation algorithm exploits Hough Circleness transformation and thresholding to find an initial mask, and Isopererimetric Distance Trees (IDT) algorithm to cut initial mask. Hessian based Frangi Vesselness filter is used for ostia points detection and small arteries segmentation. Then topological structure reconstruction is produced using thinning approach.

The result is the network of 1D vessels presented in Fig. 1. Geometrical parameters of the vessels are presented in Table 1. The values of the parameters R_k were set as follows (see Fig. 1 for vessel's notation): 20 $\frac{\text{Ba}\cdot\text{s}}{\text{cm}^3}$ for the aorta (No. 1 and 2), 7200 $\frac{\text{Ba}\cdot\text{s}}{\text{cm}^3}$ for the right coronary artery (No. 3–9), 720 $\frac{\text{Ba}\cdot\text{s}}{\text{cm}^3}$ for the left coronary artery (No. 10–18). The values of the parameters c_k were set as follows: 1050 cm/s for the aortic root (No. 1), 840 cm/s for the aorta (No. 2), 1200 cm/s for the right coronary artery (No. 3–9), 950 cm/s for the left coronary artery (No. 10–18). Parameters c_k

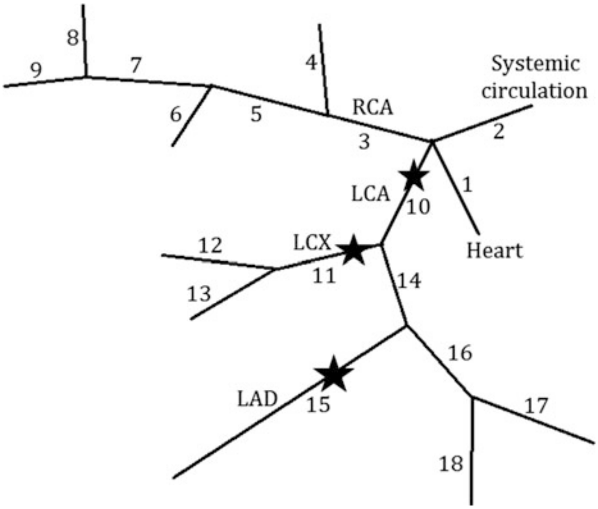


Fig. 1 The 1D reconstruction of the arterial part. Stars designate stenoses. *RCA*—right coronary artery, *LCA*—left oronary artery, *LCX*—left circumflex artery, *LAD*—left anterior descending artery

Table 1 Parameters of the arterial tree: l_k is the length, d_k is the diameter

k	l_k, cm	d_k, mm	k	l_k, cm	d_k, mm	k	l_k, cm	d_k, mm	k	l_k, cm	d_k, mm
1	5.28	21.7	6	6.75	1.52	11	6.1	3.0	16	5.4	1.91
2	60.0	25.1	7	5.01	2.50	12	2.05	1.17	17	0.38	1.01
3	2.72	3.1	8	1.27	1.19	13	1.75	1.21	18	2.62	1.19
4	1.44	1.31	9	5.65	0.157	14	1.39	3.8			
5	1.40	2.73	10	0.59	3.6	15	12.1	2.05			

were taken from clinical studies of pulse wave velocity [6]. Parameters R_k were fitted to set blood flow in coronary arteries to physiological conditions [2, 3].

The patient considered in this work was diagnosed with stenosis in three vessels (see Fig. 1): $\theta_{LCA} = 0.45$, $\theta_{LCX} = 0.2$, $\theta_{LAD} = 0.5$, where θ is lumen fraction. Stenoses were simulated by separating diseased vessel into three parts: stenosed, proximal and distal. Parameters of the stenosed part were set as $A_{0_{st}} = \theta A_0$, $R_{st} = \dfrac{l_{st}}{l\theta^2}R$, where $l_{st} = l/3$ is length of the stenosed part.

2.2 Numerical Implementation

Equations (1), (2) are closed by (3) and solved by the explicit minimum oscillating second-order grid-characteristic method from [4]. Boundary conditions are set by combining Pouseuille's pressure drop condition (5) for every vessel in junction and mass conservation. This set should be extended by compatibility conditions for hyperbolic set (1), (2) along outgoing characteristic. The result of the finite-difference discretization of the compatibility conditions of (1), (2) on the upper time layer t^{n+1} can be reduced to the form

$$u(t_{n+1}, \tilde{x}_k) = \alpha_{L,R}^{n+1} A(t_{n+1}, \tilde{x}_k) + \beta_{L,R}^{n+1}. \tag{9}$$

where $\tilde{x}_k = 0, L_k$; index L corresponds to the input of the vessel; index R corresponds to the input of the vessel. In particular, the second order approximation at the vessel's input gives

$$\begin{aligned} \alpha_L^{n+1} = w_0^n, \quad \beta_L^{n+1} = [w_0^n(\sigma_0^n(2S_1^{n+1} - \frac{1}{2}S_2^{n+1}) - S_0^n) - \\ - (\sigma_0^n(2u_1^{n+1} - \frac{1}{2}u_2^{n+1}) - u_0^n) + \tau^{n+1} f_{fr,0}^{n+1}]/(1 - \frac{3}{2}\sigma_0^n), \end{aligned} \tag{10}$$

where $\tau^{n+1} = t_{n+1} - t_n$; $\sigma = \dfrac{\tau_{n+1}}{h}\lambda_1$; $\lambda_1 = u - \sqrt{\dfrac{A}{\rho}\dfrac{\partial p}{\partial A}}$; h is the spatial step; $w = \sqrt{\dfrac{1}{\rho S}\left(\dfrac{\partial p}{\partial S}\right)}$. At the vessel's output

$$\alpha_R^{n+1} = -w_J^n, \quad \beta_R^{n+1} = [-w_J^n(\sigma_J^n(\frac{1}{2}S_{J-2}^{n+1} - 2S_{J-1}^{n+1}) - S_J^n) - \\ - (\sigma_J^n(\frac{1}{2}u_{J-2}^{n+1} - 2u_{J-1}^{n+1}) - u_J^n) + \tau^{n+1} f_{fr,0}^{n+1}]/(1 + \frac{3}{2}\sigma_J^n). \tag{11}$$

where J is the index of rightmost node.

Convergence of the numerical scheme was studied by consecutive increase the number of spatial segments by a factor of two. Each simulation was performed for a single vessel with diameter 2.1 cm, length 10 cm and $c_k = 700$ cm/s. At the input blood flow was prescribed (6); at the output the constant pressure 10 mmHg was set. Error was calculated by the Runge method as

$$E^k = \max_i \left| \frac{u_i^k - \tilde{u}_i}{\tilde{u}_i} \right|, \tag{12}$$

where u^k is the numerical solution on a grid with $10 \cdot 2^k + 1$ nodes and $\tilde{u} = u^5$ is the solution on the grid with $10 \cdot 2^5 + 1$ nodes. Comparison is shown on Fig. 2 using logarithmic scale. Numerical simulations provide the values of the order of

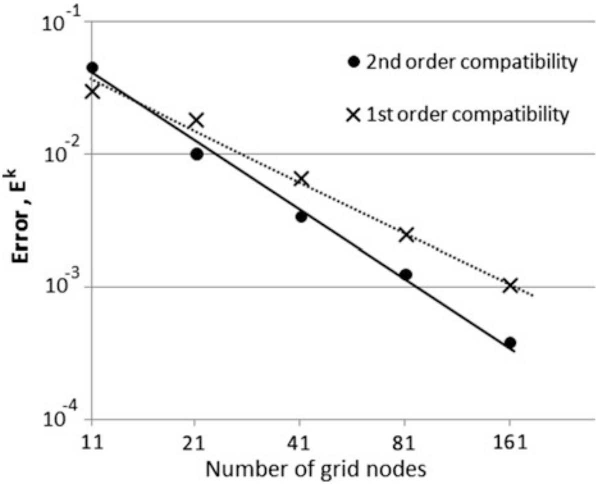


Fig. 2 Grid convergence of numerical scheme: *solid line*—2nd order approximation, *dashed line*—1st order approximation [7]

convergence 1.8 and 1.3 for the 2nd order method for internal nodes and 2nd and 1st order approximations of compatibility conditions respectively.

3 Results

Computational model from Sect. 2 was applied for simulating the change of FFR due to variability of the vessels elasticity and ARR. The FFR is calculated under the vasodilator administration as

$$FFR = \frac{\overline{P}_{dist}}{\overline{P}_{aortic}}, \qquad (13)$$

where $\overline{P}_{dist}$ is the average pressure in coronary artery distal to the stenosis, $\overline{P}_{aortic}$ is the average aortic pressure. Vasodilator administration is simulated by doubling S_0 and decreasing R by the factor of 5. This method showed good agreement with clinical results with relative error less than 17 % [2]. Two series were performed: for the stenosis in LAD of 55 % (Sect. 2.1) and 95 %.

Studying the change of the FFR due to the variability of the vessels elasticity was performed by multiplying c_k by ϵ ($0.4 \leqslant \epsilon \leqslant 2.0$). It represents typical physiological and pathological range ($\epsilon = 1$ corresponds to the reference values (see Table 1), $\epsilon < 1$ and $\epsilon > 1$ corresponds to the increased elasticity and stiffness). Substantial change of FFR up to ±5 % is observed from Fig. 3 for the $\epsilon \approx 1$. It corresponds to the variation of the c_k up to ±20 %. Both substantial stiffening and softening results in asymptotic behavior.

Studying the change of the FFR due to the variability of the ARR was performed by γ alteration ($0 \leqslant \gamma \leqslant 1$). Dramatical decrease of the FFR (up to 95 %) is observed from Fig. 4 for the values $\gamma < 0.5$. It may correspond to the high concentration of vasodilator or autoregulation failure.

In both cases the increase of the stenosis in LAD from 55 % up to 95 % results in FFR decrease up to 14 % and does not affect other vessels. It means that severe stenoses are more sensitive. We conclude that ARR strongly affects the coronary blood flow. It should be carefully included to the numerical models. Impact of the

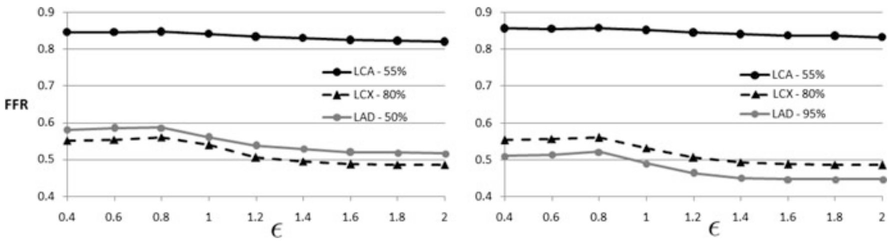


Fig. 3 The change of FFR due to the variation of elasticity

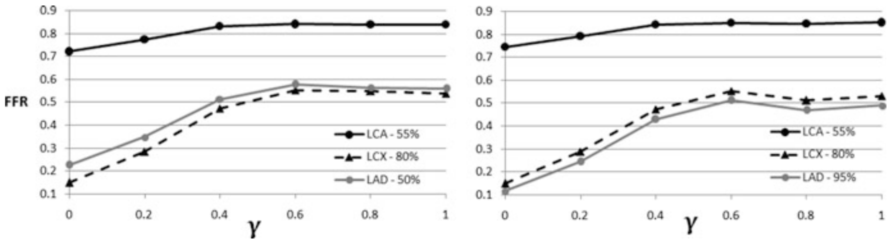


Fig. 4 The change of FFR due to the variation of ARR

elasticity variation is less recognizable. According to our study accuracy of the c_k should be less than 10 %.

4 Discussion

Our ultimate goal is to develop a computational tool for fast and robust noninvasive evaluation of FFR for individuals based on the limited patient-specific noninvasively collected data. In this work we developed numerical model based on 2nd order approximation for both internal nodes along the vessels and vessel's junctions. The numerical implementation was tested by grid refinement study. It showed better (order of 1.8) convergence rate than the model of the explicit monotone 1st order approximation (order of 1.3).

The model was applied for simulating the change of FFR due to variability of the vessels elasticity and ARR. This study was motivated by the two facts: these parameters can't be measured for individual patient in regular clinic; sensitivity of FFR to these parameters may be substantial. From numerical experiments we observe that accuracy of the elasticity estimation should be better than 10 %. Impact of ARR is up to 95 %, which means ARR itself should be thoroughly analyzed and the use of vasodilators should be carefully planned. In some numerical experiments the values of FFR were simulated under active autoregulation. It is not standard for invasive measurements during vasodilators administration in clinic. Thus we conclude (see Fig. 4), that measured FFR may be underestimated. The further joint clinical and computational research may improve personal FFR evaluation.

Acknowledgements The research was supported by Russian Science Foundation (RSF) grant 14-31-00024.

References

1. E. Boileau, P. Nithiarasu, One-dimensional modelling of the coronary circulation. application to noninvasive quantification of fractional flow reserve (FFR). Comput. Exp. Biomed. Sci.: Methods Appl. **21**, 137–155 (2015)
2. T.M. Gamilov, P.Y. Kopylov, R.A. Pryamonosov, S.S. Simakov, Virtual fractional flow reserve assessment in patient-specific coronary networks by 1D hemodynamic model. RJNAMM **30**(5), 269–276 (2015)
3. W.F. Ganong, *Review of Medical Physiology* (Appleton and Lange, Stamford, 1999)
4. K.M. Magometov, A.S. Kholodov, *Grid Characteristic Numerical Methods* (Nauka, Moscow, 1988), pp. 103–104. In Russian
5. P.D. Morris, D. Ryan, A.C. Morton et al., Virtual fractional flow reserve from coronary angiography: modeling the significance of coronary lesions. Results from the VIRTU-1 (VIRTUal fractional flow reserve from coronary angiography) study. JACC: Cardiovasc. Interv. **6**(2), 149–157 (2013)
6. R. Sala, C. Rossel, P. Encinas, P. Lahiguera, Continuum of pulse wave velocity from young elite athletes to uncontrolled older patients with resistant hypertension. J. Hypertens. **28**, 19.216 (2010)
7. S.S. Simakov, T.M. Gamilov, Y.N. Soe, Computational study of blood flow in lower extremities under intense physical load. RJNAMM **28**(5), 485–504 (2013)
8. Yu.V. Vassilevski, A.A. Danilov, T.M. Gamilov et al., Patient-specific anatomical models in human physiology. RJNAMM **30**(3), 185–201 (2015)
9. M.A. Vis, P.H. Bovendeerd, P. Sipkema, N. Westerhof, Effect of ventricular contraction, pressure, and wall stretch on vessels at different locations in the wall. Am. J. Physiol. Heart Circ. Physiol. **272**, H2963–H2975 (1997)
10. C.K. Zarins, C.A. Taylor, J.K. Min, Computed fractional flow Reserve (FFTCT) derived from coronary CT angiography. J. Cardiovasc. Trans. Res. **6**(5), 708–714 (2013)

On the Mathematical Modeling of Monocytes Transmigration

Oualid Kafi, Adélia Sequeira, and Soumaya Boujena

Abstract Monocytes play a significant role in the atherosclerosis development. During the inflammation process, monocytes that circulate in the blood stream are activated. Upon activation, they adhere to the endothelium and extravasate through the latter to migrate into the intima. In this work we are concerned with the transmigration stage. Micropipette aspiration experiments show that monocytes behave as polymeric drops during suction. In our study, the constitutive equations for Oldroyd-B fluids are used to capture the viscoelastic behavior of monocytes. We first establish and analyze a simplified mathematical model describing the coupled deformation-flow of an individual monocyte in a microchannel. Then we describe the numerical implementation of the mathematical model using the level set method and show the numerical results. Further extensions of this model are also discussed.

1 Introduction

The formation and development of atherosclerotic plaques result from a biochemical and mechanical interaction between the vessel wall and circulating blood. The crucial step in the formation of plaques is the presence of oxidized lipoproteins in the subendothelial region, which causes the local activation of the factors responsible for the infiltration of the monocytes and T lymphocytes in the vessel wall. Thus atherosclerosis is an immuno-inflammatory disease, with various components of the immune system assuming a protective or deleterious role in its development. The study of mathematical and numerical models of atherosclerosis has the potential to understand better the inflammation process and eventually to contribute to its treatment. The accumulation and continued recruitment of monocytes are associated

O. Kafi (✉) • A. Sequeira
Department of Mathematics, IST, University of Lisbon, Lisbon, Portugal

CEMAT. Av. Rovisco Pais 1, 1049-001 Lisboa, Portugal
e-mail: oualid.kafi@tecnico.ulisboa.pt; adelia.sequeira@math.ist.utl.pt

S. Boujena
FSAC, Hassan II University of Casablanca, B.P 5366, Maarif, Casablanca, Morocco
e-mail: boujena@gmail.com

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_49

with the development of vulnerable plaques [1]. Our main goal is to develop a simplified Oldroyd-B drop model for monocytes transmigration. In fact, numerical modeling of viscoelastic flows is of great importance for complex engineering applications involving food products, paints or plastics and many biological materials, including blood. Viscoelastic fluids are examples of a broader class of fluids called non-Newtonian fluids, that have the ability to store and release energy and for which the extra-stress tensor cannot be expressed as a linear, isotropic function of the components of the velocity gradient. One of the simplest nonlinear constitutive models for non-Newtonian incompressible viscous fluids is the rate type frame invariant Oldroyd-B model which can be derived from the kinetic theory of polymer dilute solutions.

A mathematical and numerical study of a simplified time-dependent viscoelastic flow is presented in [2]. The model was simplified in order to prove the global existence in time in Banach spaces, using the implicit function theorem and a maximum regularity property for a three fields Stokes problem. In [3], the authors propose a stabilized Galerkin/least squares implementation of the Oldroyd-B equations since the standard Galerkin finite element methods are prone to numerical oscillations, and solutions break down as fluid elasticity increases. The implementation of the Oldroyd-B equations in Comsol Multiphysics using this method show good agreement with results published in the literature. On the other hand, several papers address the computational modeling of flow dynamics of free-flowing, tethered (captured), rolling and adherent monocytes. We shall mention [4], where the authors investigate the effects of cell deformability and viscoelasticity on receptor-mediated leukocyte adhesion to endothelium or a ligand coated surface in a parallel plate flow chamber using computational fluid dynamics techniques. The leukocyte is modeled as a compound viscoelastic drop (a nucleus covered by a thick layer of cytoplasm). The nucleus, cytoplasm, and extracellular fluid are considered as Newtonian or viscoelastic liquids of high viscosity. The 3D numerical code is based on the Volume-of-Fluid method and the Giesekus constitutive equation is implemented in the code to capture viscoelasticity of the cytoplasm and nucleus. More recently, the authors in [5] developed a 2D mathematical model, in which leukocytes are regarded as compound viscoelastic capsules with a nucleus. The effects of several factors on flow dynamic characteristics of tethered cells, including the cell length, the inclination angle, the drag and lift forces acting on the cell were investigated. We shall also mention [6] where authors analyzed the coupled deformation-flow of individual leukocytes in microfluidic parabolic shear blood flow using a level set method.

In this paper we briefly present existence results for the solution of an Oldroyd-B drop model, since the theoretical study is not our goal. Unlike the usual formulation cited for example in [2, 4], in this work, the density is not constant. The choice of this formulation for our model is motivated by the fact that in the numerical approximation the density is represented by the level set function as it will be shown below. Next we will describe the numerical method to simulate the coupled deformation-flow in a microchannel with a perforated barrier modeling a leaky junction. The simulations are done using Comsol Multiphysics 4.3b [7].

2 The Simplified Oldroyd-B Problem

Let Ω be a bounded domain of $\mathbb{R}^N$, ($N = 2$ or 3), with boundary $\partial\Omega \in C^1(\Omega)$ (with a unit external normal $\mathbf{n}$) and $t \in (0, T)$. The unknowns of the simplified model are the velocity $\mathbf{u}$, the density ρ, the pressure γ and τ is the extra-stress tensor associated to the Oldroyd-B constitutive equations. $\mathbf{u}_0$, ρ_0, τ_0, μ^p (polymer viscosity), λ (relaxation time) and μ^s (solvent viscosity) are given. Then the coupled problem can be written as:

$$(\mathscr{P}_s)\begin{cases} \rho\dfrac{\partial \mathbf{u}}{\partial t} - 2\mu^s\,\mathbf{div}\,\epsilon(\mathbf{u}) + \nabla\gamma = \mathbf{div}\,\tau, \text{ in } \Omega \times (0, T), \\ \mathbf{div}\,\mathbf{u} = 0, \text{ in } \Omega \times (0, T), \\ \mathbf{u}(\cdot, 0) = \mathbf{u}_0, \text{ in } \Omega, \\ \mathbf{u}(x, t) = 0, \text{ on } \partial\Omega \times (0, T), \\ \tau + \lambda\left[\dfrac{\partial\tau}{\partial t} + (\mathbf{u}\cdot\nabla)\tau - \tau(\nabla\mathbf{u}) - (\nabla\mathbf{u})^T\tau\right] - 2\mu^p\epsilon(\mathbf{u}) = 0, \text{ in } \Omega \times (0, T), \\ \tau(\cdot, 0) = \tau_0, \text{ in } \Omega, \\ \dfrac{\partial\rho}{\partial t} + \mathbf{u}\cdot\nabla\rho = 0, \text{ in } \Omega \times (0, T) \\ \rho(\cdot, 0) = \rho_0, \text{ in } \Omega. \end{cases}$$

Since the mathematical analysis is not the main goal of this work, we list below a series of results [8, 9] leading to the proof of the existence of solution of problem $(\mathscr{P}_s)$ using a fixed point argument. The novelty of these results, when compared to the cited ones is that in this case the density is not constant. The usual notations for the Sobolev spaces are used below. For more details the reader can refer to [10, 11].

Lemma 1 *Let $q \in [1, +\infty]$ and $a_0 \in W^{1,q}(\Omega)$. Let $\mathbf{v} \in L^1(0, T, lip)$ (where lip is the space of bounded functions with bounded derivatives) such that* $\mathbf{div}\,\mathbf{v} = 0$ *and $\mathbf{v}\cdot\nu = 0$ on $\partial\Omega$. Then the transport problem*

$$(\mathscr{D})\begin{cases} \dfrac{\partial a}{\partial t} + \mathbf{v}\cdot\nabla a = 0, \textit{ in } \Omega \times (0, T) \\ a(\cdot, 0) = a_0, \textit{ in } \Omega. \end{cases}$$

has a unique solution such that $a \in C(0, T, W^{1,q}(\Omega))$ if $q < +\infty$ and $a \in L^\infty(0, T, W^{1,\infty}(\Omega)) \cap C(0, T, \bigcap\limits_{0<r<\infty} W^{1,r}(\Omega))$ if $q = +\infty$.

Furthermore for all $t \in [0, T]$, $\|a(t)\|_{W^{1,q}(\Omega)} \leq e^{\int_0^t \|\nabla\mathbf{v}(s)\|_{L^\infty(\Omega)}\,ds} \|a_0\|_{W^{1,q}(\Omega)}$ and if $a \in L^p(\Omega)$ for a $p \in [1, +\infty[$ then $\|a(t)\|_{L^p(\Omega)} = \|a_0\|_{L^p(\Omega)}$.

Lemma 2 *Let* $\mathbf{w} \in L^1(0,T,H^3(\Omega))$ *and* $\tau_0 \in H^2(\Omega)$ *then there exists a unique function* $\tau \in C(0,T,H^2(\Omega))$ *such that*

$$(\mathscr{E}\mathscr{S})\begin{cases}\tau + We\left[\dfrac{\partial\tau}{\partial t} + g(\tau,\nabla\mathbf{w}) + B(\mathbf{w},\theta)\right] = 2\omega\epsilon(\mathbf{w}),\ in\ \Omega\times(0,T),\\ \tau(\cdot,0) = \tau_0,\ in\ \Omega.\end{cases}$$

Moreover there exists a constant C_0 *such that*

$$\|\tau\|_{L^\infty(0,T,H^2(\Omega))} \leq (\|\tau_0\|^2_{H^2(\Omega)} + \frac{2\omega}{C_0 We})\exp\ C_0\,\|\mathbf{w}\|_{L^1(0,T,H^3(\Omega))}\,.$$

$g(\tau,\nabla u) = -\theta(\nabla u) - (\nabla u)^T\theta$, $B(u,\theta) = (u\cdot\nabla)\theta$, $\mu = \mu^s + \mu^p$, $\omega = \dfrac{\mu^p}{\mu}$, *and* $We = \dfrac{\lambda U}{L}$ *(where We is the Weissenberg number, U and L are the characteristic flow velocity and channel length, respectively).*

Let us consider the following Banach space $\mathbf{D}_{A_r}^{1-\frac{1}{p},p}(\Omega)$ that stands for a fractional domain of the Stokes operator in $L^r(\Omega)$ (the vector-fields of $\mathbf{D}_{A_r}^{1-\frac{1}{p},p}(\Omega)$ have $2-\frac{2}{p}$ derivatives in $L^r(\Omega)$, are divergence-free and vanish on $\partial\Omega$). e^{-tA_r} is the semigroup associated to A_r (Helmoltz decomposition of the operator Δ).

Lemma 3 *Let* Ω *be a* $C^{2+\epsilon}$ *bounded domain (for some* $\epsilon > 0$*),*
$1 < p,r < +\infty$ *and* $q \in (N,+\infty]$ *such that* $q \geq r$. *Let* $\mathbf{u}_0 \in \mathbf{D}_{A_r}^{1-\frac{1}{p},p}(\Omega)$ *and* $f \in L^p(0,T,L^r(\Omega))$. *Moreover, we assume that* ρ *solution of* $(\mathscr{D})$ *verifies* $\forall(x,t) \in \Omega\times(0,T)$ $0 < \tilde{\rho} \leq \rho(x,t) < \check{\rho} < +\infty$ *(*$\tilde{\rho}$ *and* $\check{\rho}$ *are positive constants) and for some value* $\beta \in (0,1]$, $\rho \in L^\infty(0,T,W^{1,q}(\Omega)) \cap C^\beta([0,T],L^\infty(\Omega))$ *then the problem*

$$(\mathscr{S})\begin{cases}\rho\dfrac{\partial\mathbf{u}}{\partial t} - \mu\Delta\mathbf{u} + \nabla\gamma = f,\ in\ \Omega\times(0,T),\\ \mathbf{div}\,\mathbf{u} = 0,\ in\ \Omega\times(0,T),\\ \mathbf{u}(\cdot,0) = \mathbf{u}_0,\ in\ \Omega,\\ \mathbf{u}(x,t) = 0,\ on\ \partial\Omega\times(0,T),\end{cases}$$

has a unique solution $(\mathbf{u},\gamma)$ *such that*
$\mathbf{u} \in C([0,T],\mathbf{D}_{A_r}^{1-\frac{1}{p},p}(\Omega)) \cap L^p(0,T,W^{2,r}(\Omega)\cap W_0^{1,r}(\Omega))$, $\gamma \in L^p(0,T,W^{1,r}(\Omega))$ *and* $\dfrac{\partial\mathbf{u}}{\partial t} \in L^r(0,T,L^p(\Omega))$,

Based on the previous results we can prove the following theorem:

Theorem 4 *Let* $1 < p < +\infty$, $q \in (N, +\infty]$ *such that* $q \geq 2$, $u_0 \in D_{A_2}^{1-\frac{1}{p},p}(\Omega)$, $\tau_0 \in H^2(\Omega)$, $\rho_0 \in W^{1,q}(\Omega)$ *and* $T > 0$ *satisfying* $\|u_0\| < 2$. *Considering the application*

$$\phi : R_T \longrightarrow \mathscr{C}([0,T], W^{3,2}(\Omega)) \times \mathscr{C}([0,T], W^{1,q}(\Omega)) \times \mathscr{C}([0,T], H^2(\Omega))$$
$$(\tilde{\mathbf{u}}, \tilde{\rho}, \tilde{\tau}) \longmapsto (\mathbf{u}, \rho, \tau)$$

where: (a) ρ *is the unique solution of* $(\mathscr{D})$ *with* $\mathbf{v} = \tilde{\mathbf{u}}$, *(b)* τ *solution of* $(\mathscr{E}S)$ *with* $\mathbf{w} = \tilde{\mathbf{u}}$ *and (c)* $\mathbf{u}$ *solution of* $(\mathscr{S})$ *with* $\rho = \tilde{\rho}$ *and* $f = \mathbf{div}(\tilde{\tau})$, *then the application* ϕ *has a fixed point for some* $0 < T^* \leq T$. *This fixed point is a solution of problem* $(\mathscr{P}_s)$.

3 Numerical Results

3.1 Setting up the Numerical Model

As a first approximation the microchannel can be considered as a rectangular geometry separated by a perforated structure, see Fig. 1. The motion of a single cell was explored in this computational domain: $12R \times 8R$. The undisturbed flow without any cell inside is considered as fully developed with a parabolic velocity profile $\mathbf{u}_0$. The fluid enters the microchannel by the left hand side, such that the velocity changes along the Y-axis as $\mathbf{u}_0 = (U\frac{y}{H}(1 - \frac{y}{H}), 0)$. Here H is the height of the channel above the structure, which is four times the radius of the cell R, and $U = 1.05 \times 10^{-3}$ (m/s) is the centerline velocity of the undisturbed parabolic flow. On the structure boundaries as well as on the edges surrounding the channel the flow satisfies no-slip boundary conditions. The boundary condition at the edge close to the inflammation zone is the null pressure. In this study, the monocyte was

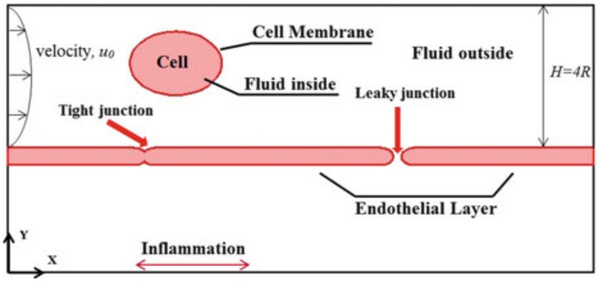


Fig. 1 Oldroyd-B drop model of individual cell. The membrane is modeled as an interface with a constant surface tension, the fluid inside the cell is modeled as an Oldroyd-B viscoelastic fluid, the fluid outside is modeled as a Newtonian fluid

assumed to be the only blood cell behaving like a liquid drop, i.e., it is deforming continuously from a spherical shape into other shapes. In the next section we give a comparison in terms of the shape deformation and volume fraction when the intracellular liquid is modeled as a homogeneous Newtonian viscous liquid drop [6], and when it is a viscoelastic drop.

3.2 Numerical Implementation

The level set method is particularly well suited for the numerical study of models with fluid-interface motion. The interface is represented implicitly and the equation of motion is numerically approximated using schemes built from hyperbolic conservation laws. The level set interface finds the fluid interface by tracing the isolines of the level set function given by a function ϕ. The equation governing the transport and reinitialization of ϕ is given by $\frac{\partial \phi}{\partial t} + \mathbf{v} \cdot \nabla\phi = \gamma_r \nabla \cdot (\varepsilon_r \nabla\phi - \phi(1-\phi)\frac{\nabla\phi}{|\nabla\phi|})$, where $\mathbf{v}$(m/s) is the velocity transporting the interface, γ_r(m/s) and ε_r(m) are reinitialization parameters, such that $\varepsilon_r = h_c/2$, with h_c the characteristic mesh size in the region passed by the interface and γ_r is the maximum velocity magnitude occurring in the model. This equation is written here in the form used by Comsol in the numerical implementation. The equation is coupled to the Navier-Stokes equations through the level set function and the velocity field. The Navier-Stokes equations read as follows

$$\rho(\frac{\partial \mathbf{v}}{\partial t} + \mathbf{v} \cdot \nabla\mathbf{v}) = -\nabla\sigma + F_{st}, \; \mathbf{div}\, \mathbf{v} = 0,$$

where the surface tension force is $F_{st} = \nabla \cdot [\sigma_{st}\{\mathbf{I} + (-\mathbf{n}\mathbf{n}^T)\}\delta]$, $\mathbf{I}$ is the identity matrix, $\mathbf{n}$ is the interface unit normal, and δ is a Dirac delta function. As already seen in Sect. 2, σ denotes the total stress tensor $\sigma(p, \mathbf{v}, \tau) = -p + 2\mu_c^s \epsilon(\mathbf{u}) + \tau$. The Oldroyd-B constitutive equation is given by: $\tau + \lambda \left[\frac{\partial \tau}{\partial t} + (\mathbf{v} \cdot \nabla)\tau - \tau(\nabla\mathbf{v}) - (\nabla\mathbf{v})^T \tau \right] - 2\mu_c^p \epsilon(\mathbf{v}) = 0$, where $\epsilon(\mathbf{v}) = \frac{1}{2}(\nabla\mathbf{v} + \nabla\mathbf{v}^T)$ is the rate of the strain tensor. If $\tau = 0$ the Oldroyd-B model will be reduced to the Newtonian model.

The solution vector contains the unknowns $[\mathbf{v}\, p\, \phi\, \tau]$. The resulting coupled system was solved using a finite element method based on Lagrange quadratic elements for the flow velocity and extra stress components and Lagrange linear elements for the pressure, with a maximum time step of 10^{-3}s. All the computations were performed with the automatic mesh generated by the direct Pardiso solver which corresponds to 14,119 degrees of freedom in the case of the Newtonian droplet model and 17,083 in the case of viscoelastic Oldroyd-B droplet model.

3.3 Results

In Fig. 2 the images show how the fluid 2 (cell) travels down through the fluid 1 (plasma) and adheres to the structure below (endothelial layer). When the cell crawls, its shape remains circular due to the surface tension and the high viscosity of the cell (parameters of monocytes used in this model are given in Table 1). When the cell reaches the hole (the leaky junction), it deforms to transmigrate in the direction of the outlet pressure free (inflammation zone). The viscosity plays an important role in the motion and deformation of the cell. A constant viscosity accelerates the motion and deformation of the cell modeled as a Newtonian liquid drop. In contrast, when the cell is modeled as an Oldroyd-B viscoelastic droplet, the viscosity is higher (see Table 1) and the cell's motion and deformation slow down significantly. The results lead us to estimate the risk associated with the development of vulnerable plaques. Hence the Newtonian droplet model overestimates the risk of plaque formation since in this case there is a faster accumulation and continued recruitment of monocytes, compared to the behavior of a cell modeled as an Oldroyd-B viscoelastic droplet.

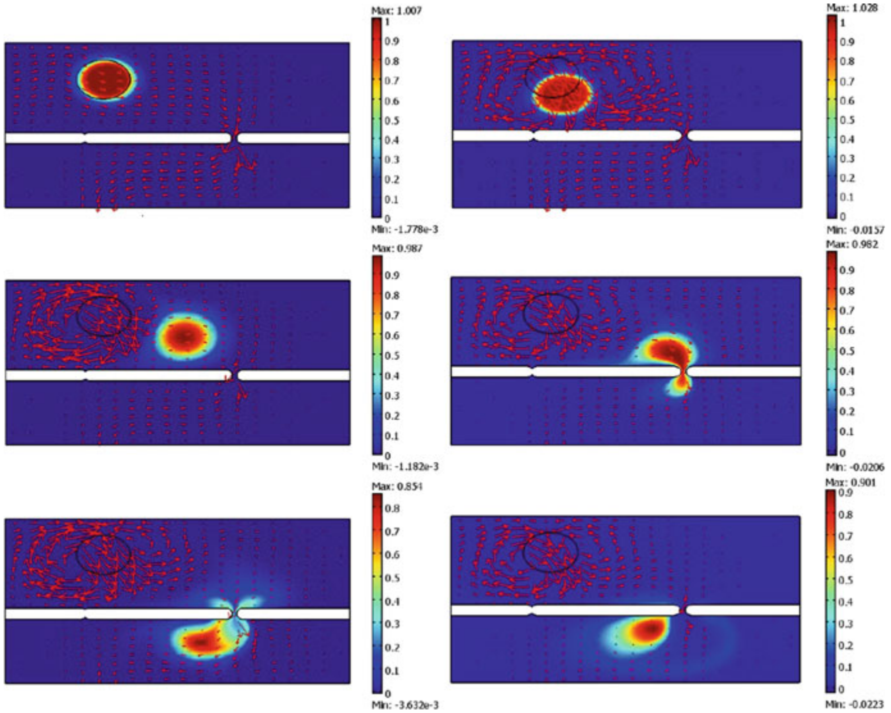


Fig. 2 Snapshots showing the volume fraction of fluid 2 (cell) and the arrow velocity field at t =0.01s, 0.2s and 0.4s (from the *top* to the *bottom*) for Oldroyd-B and Newtonian droplet models (*left* and *right*, respectively)

Table 1 Parameters used for the numerical simulations, some of them were taken from [6] and others were provided by the Institute of Molecular Medicine (IMM – FML) in Lisbon, Portugal. (Subscripts c and ec denote cell and extracellular fluid)

Parameters	Value	Description	Parameters	Value	Description
R_m	5.5 μm	Monocyte radius	R_E	35 μm	Endothelial cell radius
ρ_{ec}	1000 kg·m^{-3}	Plasma density	ρ_c	1063 kg·m^{-3}	Monocyte density
μ_{ec}	0.001 Pa·s	Plasma viscosity	μ_c^p	0.05 Pa·s	Monocyte polymer viscosity
σ_{st}	24 μN/m	Surface tension	μ_c^s	0.001 Pa·s	Monocyte solvent viscosity

4 Conclusion

We developed and analyze a simplified mathematical model describing the coupled deformation-flow of an individual monocyte in a microchannel. This model has been implemented in Comsol, using a non-conservative level set method for the simulation of the two phase flow interface. Results obtained using models for a Newtonian droplet liquid and an Oldroyd-B viscoelastic droplet have been qualitatively compared. This preliminary work describes the theory and implementation of the level set method for two fluids with different flow behaviors. Since the subendothelial layer is a conjonctive tissue, future studies should focus on developing more realistic models for this structure. However, we conclude that the results obtained so far are, qualitatively, in agreement with those observed by the experimentalists working in the field.

Acknowledgements FCT (Fundação para a Ciência e a Tecnologia, Portugal) the project EXCL/MAT-NAN/0114/2012 and the Research Center CEMAT-IST are gratefully acknowledged. The authors are also thankful to Ana Silva-Herdade from IMM – FML, for fruitful discussions and for providing some of the parameters.

References

1. K.J. Woollard, F. Geissmann, Monocytes in atherosclerosis: subsets and functions. Nat. Rev. Cardiol. **2**, 77–86 (2010)
2. A. Bonito, P. Clément, M. Picasso, Mathematical and numerical analysis of a simplified time-dependent viscoelastic flow. Numer. Math. **107**, 213–255 (2007)
3. T.J. Craven, J.M. Rees, W.B. Zimmerman, Stabilised finite element modelling of oldroyd-B viscoelastic flows, in *Excerpt from the Proceedings of the Comsol Users Conference*, Birmingham (2006)
4. D.B. Khismatullin, The cytoskeleton and deformability of white blood cells. Leukocyte rolling and adhesion. Curr. Top. Membr. Transp. Elsevier **64**, 47–111 (2009)

5. Z.Y. Luo, L. He, S.Q. Wang, S. Tasoglu, F. Xu, U Demirci, B.F. Bai, Two-dimensional numerical study of flow dynamics of a nucleated cell tethered under shear flow. Chem. Eng. Sci. **119**, 236–244 (2014)
6. Z.Y. Luo, F. Xu, T.J. lu, B.F. Bai, Direct numerical simulation of single leukocyte deformation in microchannel flow for disease diagnosis. J. Med. Syst. **35**, 869–876 (2011)
7. COMSOL Multiphysics, *User's Guide 4.3b*, Licence 17073661, 2012
8. C. Guillopé, J.C. Saut, Existence results for the flow of viscoelastic fluids with a differential constitutive law. Nonlinear Anal. **15**, 849–869 (1990)
9. R. Danchin, Density-dependent incompressible fluids in bounded domains. J. Math. Fluid Mech. **8**, 333–381 (2006)
10. J. Huang, M. Paicu, P. Zhang, Global well-posedness of incompressible inhomogeneous fluid systems with bounded density or non-lipschitz velocity. Arch. Ration. Mech. Anal. **209**, 631–682 (2013)
11. K. Zhao, Large time behavior of density-dependent incompressible navierstokes equations on bounded domains. J. Math. Fluid Mech. **14**, 471–483 (2012)

Part VIII
Computational Methods for Multi-physics Phenomena

Parallel Two-Level Overlapping Schwarz Methods in Fluid-Structure Interaction

Alexander Heinlein, Axel Klawonn, and Oliver Rheinbach

Abstract Parallel overlapping Schwarz preconditioners are considered and applied to the structural block in monolithic fluid-structure interaction (FSI). The two-level overlapping Schwarz method uses a coarse level based on energy minimizing functions. Linear elastic as well as nonlinear, anisotropic hyperelastic structural models are considered in an FSI problem of a pressure wave in a tube. Using our recent parallel implementation of a two-level overlapping Schwarz preconditioner based on the Trilinos library, the total computation time of our FSI benchmark problem was reduced by more than a factor of two compared to the algebraic one-level overlapping Schwarz method used previously. Finally, also strong scalability for our FSI problem is shown for up to 512 processor cores.

1 The Two-Level Overlapping Schwarz Preconditioner

The GDSW preconditioner [5] is a two-level additive Schwarz preconditioner

$$M_{\mathrm{GDSW}}^{-1} = \Phi \left(\Phi^T A \Phi\right)^{-1} \Phi^T + \sum_{i=1}^{N} R_i^T \tilde{A}_i^{-1} R_i, \tag{1}$$

with a special choice of energy minimizing coarse space functions Φ. The coarse space functions are discrete harmonic extensions of the restrictions of the null space of A to connected components (vertices, edges, and faces) of the interface Γ of a nonoverlapping domain decomposition. For the elasticity problems considered here, the null space is spanned by the three translations r_1, r_2, r_3 and three (linearized) rotations r_4, r_5, r_6.

A. Heinlein (✉) • A. Klawonn
Mathematisches Institut, Universität zu Köln, Weyertal 86-90, 50931 Köln, Germany
e-mail: alexander.heinlein@uni-koeln.de; axel.klawonn@uni-koeln.de

O. Rheinbach
Fakultät für Mathematik und Informatik, Institut für Numerische Mathematik und Optimierung, Technische Universität Bergakademie Freiberg, Akademiestr. 6, 09596 Freiberg, Germany
e-mail: oliver.rheinbach@math.tu-freiberg.de

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_50

Let $I = \Omega \setminus \Gamma$ be the set of degrees of freedom (d.o.f.) in the interior of a subdomain. Then the basis functions of the GDSW coarse space are defined by

$$\Phi = \begin{bmatrix} -A_{II}^{-1} A_{\Gamma I}^{T} \Phi_{\Gamma} \\ \Phi_{\Gamma} \end{bmatrix} = \begin{bmatrix} \Phi_I \\ \Phi_{\Gamma} \end{bmatrix}, \tag{2}$$

where $\Phi_\Gamma = \left[R_{\Gamma 1}^T G_{\Gamma 1} \dots R_{\Gamma M}^T G_{\Gamma N} \right]$, and $R_{\Gamma j}$ is the restriction from Γ onto Γ_j, the j-th interface component. The matrices $G_{\Gamma j}$ are chosen such that their columns form a basis of the restriction of the matrix G to the indices corresponding to Γ_j. For GDSW, the condition number bound

$$\kappa \left(M_{GDSW}^{-1} K \right) \leq C \left(1 + \frac{H}{\delta} \right) \left(1 + \log \left(\frac{H}{h} \right) \right)$$

was shown in [5]. Here, H and h are the typical subdomain and finite element diameters and δ denotes the overlap.

We have implemented a parallel GDSW preconditioner based on Trilinos [8] and report on parallel scalability for model problems in [7]. Currently, we use UMFPACK 5.3.0 to solve the problems on the subdomains (from algebraic or geometric overlapping Schwarz) and MUMPS 4.10.0 in MPI-parallel mode for the coarse problem. From our experience, on modern AMD or Intel processors, UMFPACK is often not slower than MUMPS if the matrices are small although Umfpack often uses more memory.

We use a one-to-one correspondence of subdomains to cores.

2 Monolithic Fluid-Structure Interaction

We use the software environment from [3], i.e., we use the LifeV software library 3.6.2 coupled to FEAP 8.2. As opposed to [3], where a convective explicit (CE) approach was used for the fluid, we now use a fully implicit scheme, and the linearized systems are now preconditioned using a FaCSI preconditioner applying a SIMPLE preconditioner for the fluid; see [4].

2.1 Model Description

The fluid-structure interaction (FSI) problem consists of the fluid problem

$$\begin{cases} \rho_f \left(\left. \frac{\partial \mathbf{u}}{\partial t} \right|_{\mathbf{X}} + ((\mathbf{u} - \mathbf{w}) \cdot \nabla)\mathbf{u} \right) - \nabla \cdot \boldsymbol{\sigma}_f(\mathbf{u}, p) = 0 \text{ in } \Omega_t^f \times (0, T], \\ \nabla \cdot \mathbf{u} = 0 \qquad \text{in } \Omega_t^f \times (0, T], \end{cases} \tag{3}$$

which corresponds to the incompressible Navier-Stokes equations in Arbitrary Lagrangian Eulerian (ALE) formulation, cf. [6], and the structural problem

$$\rho_s \frac{\partial^2 \mathbf{d}_s}{\partial t^2} - \nabla \cdot (\mathbf{FS}) = 0 \quad \text{in } \Omega^s \times (0, T]. \tag{4}$$

Here, Ω_t^f and Ω^f are the fluid domain in the actual and the reference configuration, respectively, Ω_s is the structural reference configuration, and $\Gamma = \partial\Omega^f \cap \partial\Omega^s$ is the FSI interface. In (3), $\frac{\partial}{\partial t}|_{\mathbf{X}} = \frac{\partial}{\partial t} + \mathbf{w} \cdot \nabla$ is the ALE derivative and $\mathbf{X}$ corresponds to the fluid coordinates in reference configuration, ρ_f denotes the density of the fluid, $\mathbf{u}$ and p are the velocity and pressure, respectively, $\mathbf{w} = \frac{\partial \mathbf{d}_f}{\partial t}|_{\mathbf{X}}$ is the velocity and $\mathbf{d}_f$ the displacement of the computational fluid domain, and $\boldsymbol{\sigma}_f(\mathbf{u}, p)$ is the Cauchy stress tensor. In (4), $\mathbf{d}_s$ is the displacement of the structure, ρ_s is the density of the structure, and $\mathbf{FS}$ are the first Piola-Kirchhoff stresses.

The ALE mapping $\mathscr{A}_t = id + \mathbf{d}_f$ is obtained by solving an additional geometry problem

$$\begin{cases} -\Delta \mathbf{d}_f = \mathbf{0} \text{ in } \Omega^f, \\ \mathbf{d}_f = \mathbf{d}_s \qquad \text{on } \Gamma, \\ \mathbf{d}_f \cdot \mathbf{n}_f = 0 \text{ on } \partial\Omega^f \backslash \Gamma, \end{cases}$$

i.e., by means of a discrete harmonic extension. The fluid, structural, and geometry problems are coupled by the geometric adherence (5), the continuity of the velocities (6), and the continuity of the stresses (7) on Γ,

$$\mathbf{d}_f = \mathbf{d}_s, \tag{5}$$

$$\frac{\partial \mathbf{d}_s}{\partial t} = \mathbf{u} \circ \mathscr{A}_t, \tag{6}$$

$$(\det[\mathbf{F}])^{-1} \mathbf{F}^{-T} \boldsymbol{\sigma}_f \, \mathbf{n}_f \circ \mathscr{A}_t + (\mathbf{FS}) \, \mathbf{n}_s = 0. \tag{7}$$

Here, $\mathbf{n}_f$ and $\mathbf{n}_s$ are the outer normal vectors of the fluid and the structural domain, respectively, and $\mathbf{F}$ is the deformation gradient.

2.2 Monolithic Coupling in FSI

We use finite differences, in a fully implicit scheme, for the approximation of the time derivatives of both the fluid and the structure equations. We use piecewise quadratic (P2) finite elements for the structure and geometry problems and P2–P1 mixed finite elements for the fluid, using conforming meshes at the FSI interface.

The monolithic approach leads to a single nonlinear system containing the fluid (F), the structure (S), the geometry (H) problem, and the coupling conditions,

$$\begin{pmatrix} F(\mathbf{u}_f^{n+1}, p^{n+1}, \mathbf{d}_f^{n+1}) + 0 & + C_1^T \boldsymbol{\lambda}^{n+1} + 0 \\ 0 \quad + S(\mathbf{d}_s^{n+1}) & + C_3^T \boldsymbol{\lambda}^{n+1} + 0 \\ C_1 \mathbf{u}_f^{n+1} \quad + C_2 \mathbf{d}_s^{n+1} & + 0 \quad + 0 \\ 0 \quad + C_4 \mathbf{d}_s^{n+1} & + 0 \quad + H \mathbf{d}_f^{n+1} \end{pmatrix} = \begin{pmatrix} \mathbf{b}_f \\ \mathbf{b}_s \\ C_2 \mathbf{d}_s^n \\ 0 \end{pmatrix}. \tag{8}$$

Here, $\boldsymbol{\lambda}$ is the vector of Lagrange multipliers. For conforming meshes, we have $C_1|_\Gamma = I|_\Gamma$, $C_3|_\Gamma = -I|_\Gamma$, $C_2|_\Gamma = 1/\Delta t\, C_3$, $C_4|_\Gamma = I|_\Gamma$, where $I|_\Gamma$ is the identity matrix defined on the interface Γ.

2.3 Linearization and Parallel Monolithic Preconditioner

As in [3], we solve the nonlinear monolithic FSI problem (8) using an inexact Newton method, i.e., the Newton equation is solved iteratively only up to a given tolerance. The corresponding tangent J_M, associated with (8), reads

$$J_\mathrm{M} = \begin{pmatrix} D_{(\mathbf{u}_f,\mathbf{p})}F & 0 & C_1^T & D_{\mathbf{d}_f}F \\ 0 & D_{\mathbf{d}_s}S & C_3^T & 0 \\ C_1 & C_2 & 0 & 0 \\ 0 & C_4 & 0 & H \end{pmatrix} \approx \begin{pmatrix} D_{(\mathbf{u}_f,\mathbf{p})}F & 0 & C_1^T & D_{\mathbf{d}_f}F \\ 0 & D_{\mathbf{d}_s}S & \mathbf{0} & 0 \\ C_1 & C_2 & 0 & 0 \\ 0 & C_4 & 0 & H \end{pmatrix} =: P_{DN}. \tag{9}$$

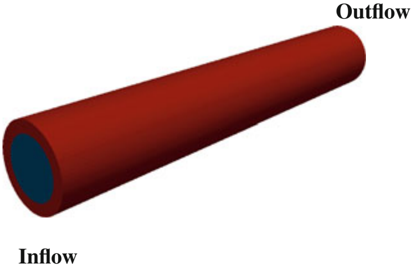
Here, $D_{(\mathbf{u}_f,\mathbf{p})}F$ denotes the linearization of the fluid operator, $D_{\mathbf{d}_f}F$ the shape derivatives, and $D_{\mathbf{d}_s}S$ the linearization of the structural operator.

We solve the linearized system using a GMRES iteration with the FaCSI preconditioner [4], which is based on a factorization of the matrix P_{DN}. The fluid block is treated further by static condensation of the interface degrees of freedom and the use of a SIMPLE preconditioner for the fluid block; see [4]. The inverses appearing in the application of the FaCSI preconditioner are replaced by algebraic (for geometry, fluid, or structure, separately) and geometric one-level overlapping Schwarz preconditioners (for the structure) or the GDSW preconditioner (for the structure).

3 Numerical Results for Fluid-Structure Interaction

We consider our FSI problem while applying different preconditioners to the structural block – but without replacing the IFPACK preconditioners for the fluid and geometry blocks. We then report on the resulting performance of the full monolithic FSI simulation.

The default preconditioner for the structural block is IFPACK, a parallel algebraic overlapping Schwarz preconditioner from Trilinos [8]. The systems from LifeV use block coordinate numbering, i.e., all x variable first. Our parallel preconditioner has two potential advantages over IFPACK: it uses a geometric overlap, and it can use a coarse space, for better robustness and improved numerical scalability. We consider three different meshes; cf. Fig. 1 and Table 1. We apply zero-displacement Dirichlet boundary conditions to the structure at the inlet and the outlet. A *pressure wave in a tube* (see Fig. 2) is created by applying a constant normal stress $\sigma \cdot n = 1.33\,\text{kPa}$ at the inflow for $t \leq 0.003$s. We consider three different material models for the wall, i.e., a linear elastic (LE), a Neo-Hookean (NH), and a realistic anisotropic nonlinear material model (Ψ_A), cf. [1, 2], which we have already considered in FSI for arterial walls in [3]. For linear elasticity, we use $E = 400\,\text{kPa}$ and $\nu = 0.3$, for Neo-Hooke,



Mesh #1:	Interior radius of the structure	0.15cm
	Outer radius of the structure	0.21cm
	Length	2.5cm
Mesh #2:	Interior radius of the structure	0.08cm
	Outer radius of the structure	0.1cm
	Length	5cm
Mesh #3:	Interior radius of the structure	0.08cm
	Outer radius of the structure	0.11cm
	Length	10cm

Fig. 1 Geometry of the FSI problem. The number of d.o.f. is almost identical for all geometries and well-balanced between fluid (F) and structure (S), cf. Table 1

Table 1 Number of degrees of freedom of the different meshes

Mesh	Velocity (F)	Pressure (F)	Displacement (S)	Displacement (G)
#1	393,903	17,261	379,080	393,903
#2	401,763	17,775	373,032	401,763
#3	376,623	17,352	346,320	376,623

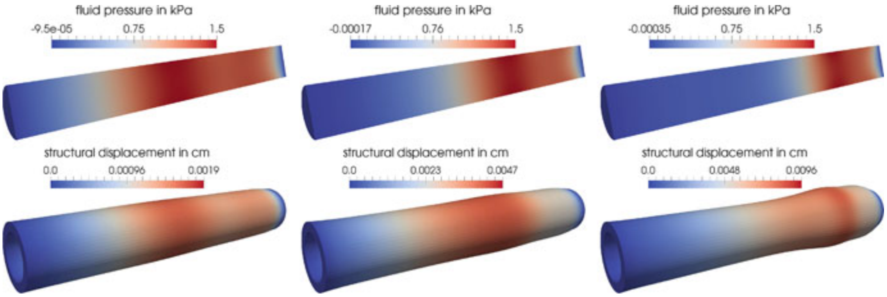


Fig. 2 Fluid pressure (*top*) and structural deformation (*bottom*) for the linear elastic (*left*), the Neo-Hookean (*middle*), and the Ψ_A (*right*) material model at $t = 0.003$s. The structural displacement is magnified by a factor of 10. The figure also illustrates the significantly different behavior for the material models

$\mu = 77.2\,\text{kPa}$ and $\kappa = 3833\,\text{kPa}$, and for the Ψ_A model, we use the parameters from [2, (Ψ_A Set 2)].

3.1 Time to Solution Using Different Preconditioners for the Structure Block

We perform simulations of the *pressure wave in a tube* for a total simulation time $T = 0.01$s. We use a time step $\Delta t = 0.0001$s, 0.0002s, 0.0004s, or 0.0005s, i.e., we solve 100, 50, 25 or 20 monolithic nonlinear systems. The nonlinear problems are solved using, on average, 5.1 (LE 0.0001s), 5.6 (NH 0.0001s), 6.6 (Ψ_A 0.0001s), 6.1 (LE 0.0002s), 6.3 (NH 0.0002s), 7.9 (Ψ_A 0.0002s), 7.4 (LE 0.0004s), 7.9 (NH 0.0004s), 11.9 (Ψ_A 0.0004s), 8.4 (LE 0.0005s), or 9.5 (NH 0.0005s) Newton iterations, and a preconditioned GMRES iteration is used to solve the linearized monolithic systems in each Newton step. Our stopping criterion for Newton is a mixed criterion with a relative and absolute tolerance of 10^{-8} , i.e., the Newton iteration is stopped when $\min\{\|r_n\|_\infty, \|r_n\|_\infty/\|r_0\|_\infty\} < 10^{-8}$, and for GMRES a relative tolerance of 10^{-6} , i.e., the GMRES iteration is stopped when $\|r_n\|_2/\|r_0\|_2 < 10^{-6}$. With $\|\cdot\|_\infty$ and $\|\cdot\|_2$ we refer to the corresponding vector norms. The computations were performed on a Cray XT6 (Universität Duisburg-Essen). We compare the number of iterations and the computing times in Table 2;

Table 2 Average computing time per time step (in minutes) and average number of GMRES iterations per Newton step for the *pressure wave in a tube* problem; see Fig. 3 for the total runtimes. Linear elasticity (LE), Neo-Hooke (NH), and a nonlinear, anisotropic hyperelastic material law to model an arterial wall (Ψ_A); see also Fig. 2. The time step is Δt and the final simulation time is $T = 0.01$s. We compare IFPACK with the one-level overlapping Schwarz preconditioner (OS1) and the GDSW preconditioner with and without rotations (GDSW/GDSW-nr) on 128 cores of a Cray XT6m. No convergence for Ψ_A and $\Delta t = 0.0005$. Best numbers in **bold face**

		IFPACK		One-level Schwarz (OS1)		GDSW w/o rot. (GDSW-nr)		GDSW	
Δt	Struct.	Time	GMRES iter	Time	GMRES iter	Time	GMRES iter	Time	GMRES iter
0.0001s	LE	**5.0 m**	53.4	5.1 m	**50.8**	5.4 m	51.8	5.3 m	**50.8**
	NH	8.6 m	89.8	**6.8 m**	59.3	7.1 m	55.0 m	7.0 m	**52.7**
	Ψ_A	19.7 m	214.7	**9.9 m**	82.0	10.5 m	81.0	10.6 m	**79.1**
0.0002s	LE	8.9 m	95.8	7.8 m	74.5	7.0 m	60.7	**6.8 m**	**58.0**
	NH	14.2 m	152.4	9.8 m	87.5	9.6 m	77.2	**9.0 m**	**66.0**
	Ψ_A	33.3 m	316.7	**13.2 m**	96.9	13.8 m	94.1	13.9 m	**90.7**
0.0004s	LE	15.3 m	147.2	14.1 m	124.5	10.9 m	84.4	**9.6 m**	**71.9**
	NH	24.7 m	226.5	17.8 m	145.7	16.2 m	117.9	**13.6 m**	**88.4**
	Ψ_A	63.0 m	399.9	27.0 m	145.4	27.1 m	135.5	**23.5 m**	**108.5**
0.0005s	LE	19.4 m	169.0	17.7 m	142.0	13.0 m	93.7	**11.3 m**	**76.3**
	NH	33.5 m	261.5	24.2 m	171.0	20.9 m	133.2	**17.1 m**	**96.1**

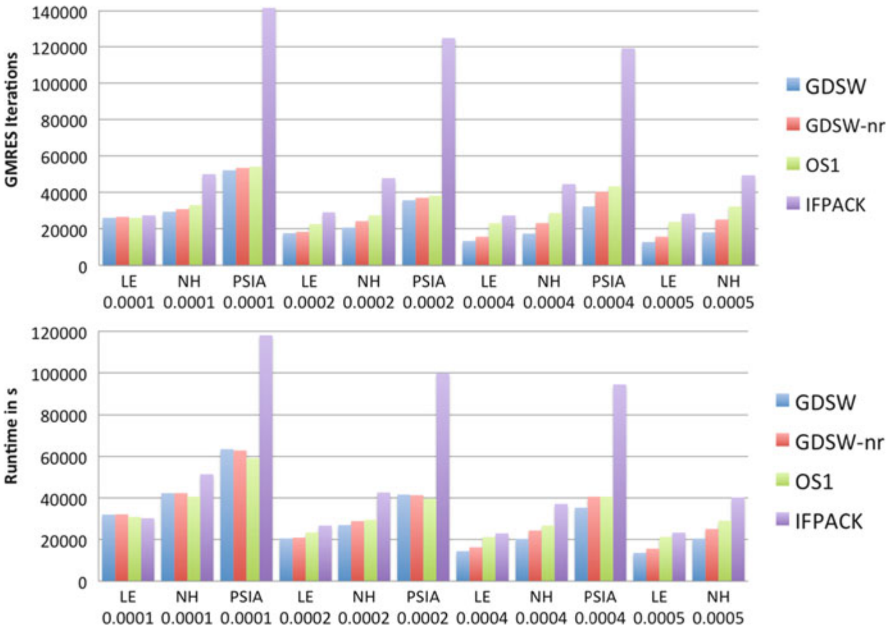


Fig. 3 Total number of GMRES iterations (*top*) and total runtime (*bottom*) for the *pressure wave in a tube* FSI problem using Mesh #1 and 128 cores; see also Table 2. We use different preconditioners for the structure block. "OS1" is the one-level Schwarz preconditioner, "GDSW-nr" is the GDSW preconditioner without rotations, and "GDSW" is the GDSW preconditioner with full coarse space

see also Fig. 3. When using our preconditioner, we consider three cases: only using the first level (OS1), using the first and coarse level but neglecting the rotations (r_4, r_5, r_6) when constructing the coarse level (GDSW-nr), and using the full GDSW preconditioner (GDSW), i.e., with first and coarse level. In the case where rotations are neglected (GDSW-nr) no geometric information is needed for the construction of the coarse problem. For all overlapping Schwarz preconditioners, including IFPACK, we specify an overlap of $\delta = 2h$. We perform the comparison using Mesh #1 and 128 cores.

In Table 2, for a small time step, all preconditioners show a very similar performance with respect to the number of GMRES iteration as well as the timings. However, for a larger time step, where the weight in front of the mass matrix is small, the number of iterations and the timings for IFPACK quickly deteriorate. The other methods, which use a geometric overlap, show a better performance. The use of a coarse space gives further improvements: for the largest time steps the GDSW preconditioner is the fastest method. Neglecting the rotations in the GDSW preconditioner (GDSW-nr), which makes the preconditioner more algebraic, yields a number of iterations which falls between the one-level preconditioner and the GDSW preconditioner with the full coarse space. For our experiments, it thus seems that for GDSW-nr it is not easy to amortize the cost for the coarse level compared to OS1.

The use of the GDSW preconditioner with the full coarse space, however, can be recommended as a new default. For smaller time steps the performance of all preconditioners is similar and for larger time steps it is clearly the fastest option: for the most challenging structural model (Ψ_A) combined with the largest time step the monolithic FSI simulation is more than 2.5 times faster when using the GDSW preconditioner instead of IFPACK; see Table 2. This is especially remarkable since, in our monolithic preconditioner, we have only exchanged the preconditioner for the structural block whereas the timings are for the complete FSI simulation.

3.2 Strong Scaling for the Fluid-Structure Interaction Problem

In Figs. 4 and 5, we present strong parallel scaling results for the first time step for the *pressure wave in a tube* using $\Delta t = 0.0001$s and $\Delta t = 0.0002$s, respectively, for a linear elastic tube. The computations were performed on the JUQUEEN supercomputer (JSC Jülich, Germany). For the structure, we have used our new default preconditioner, i.e., the GDSW preconditioner including rotations, with overlaps of $\delta = 1h$ and $\delta = 2h$. For the fluid and the geometry blocks, we have used the IFPACK preconditioner with overlap $\delta = 2h$. We present the GMRES iterations per Newton step and the total runtime for a time step. The timings are for the first time step of the fully coupled FSI simulation.

For all cases, we observe good scalability results but also a significant influence of the geometry on the performance: the properties of the domain decompositions result in different numbers of GMRES iterations. The simulation is stopped before the wave has reached the outflow. Therefore, reflections at the outflow are not relevant, here. The scaling is slightly worse for a time step of 0.0002s, which is

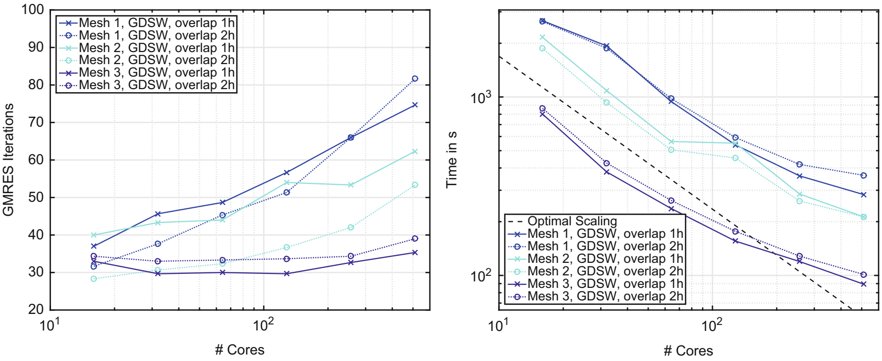


Fig. 4 Strong scaling (16 to 512 cores) for FSI using linear elasticity and $\Delta t = 0.0001$s. The computing time for one time step is shown. Always 3 Newton steps

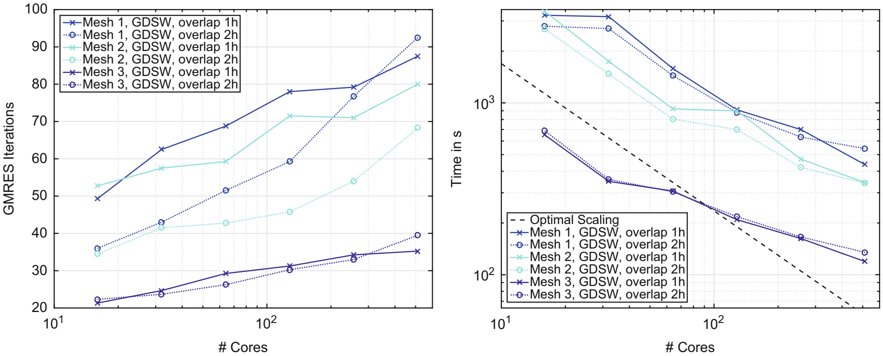


Fig. 5 Same as Fig. 4 but using $\Delta t = 0.0002$s. Newton steps vary from 3 to 5

partially a result of a higher number of Newton iterations. For Mesh #3, we observe the lowest number of iterations, the best numerical scalability, the lowest computing times, and the best parallel scalability.

Acknowledgements The authors gratefully acknowledge the Gauss Centre for Supercomputing (GCS) for providing computing time through the John von Neumann Institute for Computing (NIC) on the GCS share of the supercomputer JUQUEEN at Jülich Supercomputing Centre (JSC). The authors also gratefully acknowledge the Cray XT6m at Universität Duisburg-Essen and the financial support by the German Science Foundation (DFG), project no. KL2094/3-1 and RH122/4-1.

References

1. D. Balzani, P. Neff, J. Schröder, G.A. Holzapfel, A polyconvex framework for soft biological tissues. Adjustment to experimental data. Internat. J. Solids Struct. **43**(20), 6052–6070 (2006)
2. D. Brands, A. Klawonn, O. Rheinbach, J. Schröder, Modelling and convergence in arterial wall simulations using a parallel FETI solution strategy. Comput. Methods Biomech. Biomed. Eng. **11**, 569–583 (2008)
3. D. Balzani, S. Deparis, S. Fausten, D. Forti, A. Heinlein, A. Klawonn, A. Quarteroni, O. Rheinbach, and J. Schröder, Numerical modeling of fluid-structure interaction in arteries with anisotropic polyconvex hyperelastic and anisotropic viscoelastic material models at finite strains. Int. J. Numer. Methods Biomed. Eng. (2015). http://dx.doi.org/10.1002/cnm.2756
4. S. Deparis, D. Forti, G. Grandperrin, A. Quarteroni, FaCSI: a block parallel preconditioner for fluid-structure interaction in hemodynamics. Technical report 13, 2015
5. C.R. Dohrmann, A. Klawonn, O.B. Widlund, Domain decomposition for less regular subdomains: overlapping Schwarz in two dimensions. SIAM J. Numer. Anal. **46**(4), 2153–2168 (2008)
6. L. Formaggia, A. Quarteroni, A. Veneziani, *Cardiovascular Mathematics*, vol. 1 (Springer, Milan/New York, 2009)
7. A. Heinlein, A. Klawonn, O. Rheinbach, Parallel overlapping Schwarz with an energy-minimizing coarse space, 2015, in *To be Submitted to the Proceedings of the 23rd International Conference on Domain Decomposition Methods*. Lecture Notes in Computational Science and Engineering, TUBAF Preprint 17/2015: http://tu-freiberg.de/fakult1/forschung/preprints

8. M.A. Heroux, R.A. Bartlett, V.E. Howle, R.J. Hoekstra, J.J. Hu, T.G. Kolda, R.B. Lehoucq, K.R. Long, R.P. Pawlowski, E.T. Phipps, A.G. Salinger, H.K. Thornquist, R.S. Tuminaro, J.M. Willenbring, A. Williams, K.S. Stanley, An overview of the Trilinos project. ACM Trans. Math. Softw. **31**(3), 397–423 (2005)

Finite Volume Scheme for Modeling of NAPL Vapor Transport in Air

Ondřej Pártl, Michal Beneš, and Peter Frolkovič

Abstract We present a mathematical and numerical model for non-isothermal, compressible flow of a mixture of two ideal gases subject to gravity. This flow is described by the balance equations for mass, momentum and energy that are solved numerically by the scheme based on the method of lines. The spatial discretization is carried out by means of the finite volume method, where the staggered arrangement of variables is employed. The time integration is realized by the Runge-Kutta-Merson method. The article also contains test results obtained by the presented numerical scheme.

1 Introduction

Detailed description of compositional flow through porous media and free space above it is a part of the research carried out within the context of environment protection, new energy resources and climate change [3, 5]. The subject of our research is the NAPL (Non-Aqueous Phase Liquids) vapor transport driven by air flow in porous medium and above its surface. Our aim is to describe such a flow and to discuss its nature. In this paper, we present the part of our model describing the non-isothermal, compressible, free flow above the porous medium in which we also include gravity effects. Hence our model differs from the other models for multicomponent free flow because, to the best of our knowledge, research in this field concentrates on incompressible flows (e.g., [3, 8, 9]) or compressible flows without gravity effects (e.g., [7, 10]). These models also differ in the complexity of

O. Pártl (✉) • M. Beneš
Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University in Prague,
Trojanova 13, 120 00 Prague 2, Czech Republic
e-mail: partlond@fjfi.cvut.cz; michal.benes@fjfi.cvut.cz

P. Frolkovič
Faculty of Civil Engineering, Slovak University of Technology, Radlinského 11, 810 05
Bratislava, Slovakia
e-mail: peter.frolkovic@stuba.sk

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_51

the interaction among the species. We also present results of a numerical test carried out by our scheme.

2 Mathematical Model

From the theory of gas mixtures described in [2, 4, 6], it follows that a mixture of two polyatomic ideal gases (the first one will be referred to as ‘gas’ and the second one as ‘NAPL vapor’) can be described by the following conservation laws:

$$\partial\rho/\partial t + \nabla\cdot(\rho\boldsymbol{v}) = 0 \ \text{(conservation of mass)}, \tag{1}$$

$$\partial\rho_n/\partial t + \nabla\cdot[\rho_n(\boldsymbol{v}+\boldsymbol{V_n})] = 0 \ \text{(conservation of NAPL vapor mass)}, \tag{2}$$

$$\partial(\rho\boldsymbol{v})/\partial t + \nabla\cdot(\boldsymbol{P}+\rho\boldsymbol{v}\otimes\boldsymbol{v}) = \rho\boldsymbol{g} \ \text{(conservation of momentum)}, \tag{3}$$

$$\partial(\rho e)/\partial t + \nabla\cdot(\boldsymbol{Q}+\rho e\boldsymbol{v}+\boldsymbol{P}\cdot\boldsymbol{v}) = \rho\boldsymbol{g}\cdot\boldsymbol{v} \ \text{(conservation of energy)}. \tag{4}$$

In this paper, vectors and matrices are printed in the bold font, whereas their components are in the non-bold font, i.e., $\boldsymbol{v} = (v_1, v_2)$. The quantities which refer to the gas or NAPL vapor are denoted by the subscript g and n, respectively; the quantities without subscript refer to the whole mixture. ρ [kg $\cdot$ m^{-3}] represents the density, t [s] the time, $\boldsymbol{v}$ [m $\cdot$ s^{-1}] the velocity, ρ_i [kg $\cdot$ m^{-3}] the partial density of the component i ($\sum_{i\in\{n,g\}}\rho_i = \rho$), $\boldsymbol{V}_i$ [m $\cdot$ s^{-1}] the diffusion velocity of the component i, $\boldsymbol{P}$ [Pa] the pressure tensor and $\boldsymbol{g}$ [m $\cdot$ s^{-2}] the gravitational acceleration vector. The symbol $\otimes$ stands for the tensor product, e [m^2 $\cdot$ s^{-2}] is the specific energy and $\boldsymbol{Q}$ [kg $\cdot$ s^{-3}] the heat flow vector. The fluxes $\boldsymbol{V}_i$, $\boldsymbol{P}$ and $\boldsymbol{Q}$ are defined as

$$\boldsymbol{V}_i = -\textstyle\sum_{j\in\{n,g\}} D_{i,j}\left(\boldsymbol{d}_j + k_{Tj}\left(\nabla T/T\right)\right),\ i = g, n, \tag{5}$$

$$\boldsymbol{P} = p\boldsymbol{I} - 2\mu\boldsymbol{S}, \tag{6}$$

$$\boldsymbol{Q} = -\lambda\nabla T + p\textstyle\sum_{i\in\{n,g\}}\left(k_{Ti} + \frac{\kappa}{\kappa-1}\frac{p_i}{p}\right)\boldsymbol{V_i}, \tag{7}$$

where $\boldsymbol{d}_i$ [m^{-1}] denotes the diffusion driving force defined by

$$\boldsymbol{d}_i = \nabla\left(p_i/p\right) + \left(p_i/p - X_i\right)\left(\nabla p/p\right), \tag{8}$$

where p_i [Pa] and X_i [$-$] stand for the partial pressure and mass fraction of the component i ($\sum_{i\in\{n,g\}} p_i = p$, $\sum_{i\in\{n,g\}} X_i = 1$), respectively, and p [Pa] is the pressure. $D_{i,j}$ [m^2$\cdot$s^{-1}] is the multicomponent diffusion coefficient, where $D_{i,j} = D_{j,i}$ and $D_{i,i} = -\left(\rho_j/\rho_i\right)D_{j,i}$ if $\rho_i \neq 0$; otherwise, $D_{i,i}$ is not needed. k_{Ti} [$-$] is the thermal diffusion ratio, $k_{Tn} = -k_{Tg}$, μ [kg $\cdot$ m^{-1} $\cdot$ s^{-1}] denotes the dynamic viscosity, $\boldsymbol{S}$ [s^{-1}]

the rate-of-shear tensor defined by

$$\boldsymbol{S}_{i,j} = (1/2)\left(\partial v_j/\partial x_i + \partial v_i/\partial x_j\right) - (1/3)\,\nabla \cdot \boldsymbol{v}\delta_{i,j}, \tag{9}$$

where x_i [m], $i = 1, 2$, are spatial coordinates, and $\delta_{i,j}$ is the Kronecker delta. λ [$\mathrm{kg} \cdot \mathrm{m} \cdot \mathrm{K}^{-1} \cdot \mathrm{s}^{-3}$] denotes the thermal conductivity coefficient, T [K] the thermodynamic temperature and $\kappa = \frac{c_p}{c_v}$ [$-$] the ratio of specific heats, where c_p [$\mathrm{J} \cdot \mathrm{kg}^{-1} \cdot \mathrm{K}^{-1}$] and c_V [$\mathrm{J} \cdot \mathrm{kg}^{-1} \cdot \mathrm{K}^{-1}$] are the specific heat at constant pressure and volume, respectively.

System (1), (2), (3), and (4) is supplemented by the formula $\rho e = c_V \rho T + \frac{1}{2}\rho \boldsymbol{v}^2$ and by the ideal gas equation of state $\rho = p\overline{M}/(RT)$, where R [$\mathrm{J} \cdot \mathrm{K}^{-1} \cdot \mathrm{mol}^{-1}$] stands for the gas constant, and $\overline{M}$ [$\mathrm{kg} \cdot \mathrm{mol}^{-1}$] is the molar mass defined by $\overline{M} = \left(\sum_{i \in \{n,g\}} X_i/M_i\right)^{-1}$, where M_i [$\mathrm{kg} \cdot \mathrm{mol}^{-1}$] is the molar mass of the component i. If we combine the previous equations with the Mayer relation $\overline{M}(c_p - c_V) = R$, we get the formula which relates the energy to the pressure $p = (\kappa - 1)\left(\rho e - \frac{1}{2}\boldsymbol{v}^2 \rho\right)$.

Note that for $\rho_n = 0$, the governing equations reduce to the compressible Navier-Stokes equations and the corresponding energy balance equation.

The previous system is solved in a rectangular domain $\Omega \subset \mathbb{R}^2$ and on a time interval $[t_{\mathrm{ini}}, t_{\mathrm{fin}}]$, where the initial conditions are

$$\rho(t_{\mathrm{ini}}, \boldsymbol{x}) = \rho_{\mathrm{ini}}(\boldsymbol{x}), \qquad \rho_n(t_{\mathrm{ini}}, \boldsymbol{x}) = \rho_{n,\mathrm{ini}}(\boldsymbol{x}), \tag{10}$$

$$T(t_{\mathrm{ini}}, \boldsymbol{x}) = T_{\mathrm{ini}}(\boldsymbol{x}), \qquad \boldsymbol{v}(t_{\mathrm{ini}}, \boldsymbol{x}) = \boldsymbol{v}_{\mathrm{ini}}(\boldsymbol{x}) \tag{11}$$

for $\boldsymbol{x} \in \overline{\Omega}$. The boundary conditions will be discussed in Sect. 4.

3 Numerical Solution

The aforementioned mathematical problem is solved by means of the method of lines. The spatial discretization is carried out by the finite volume method, where the staggered arrangement of the variables is used [12]. For the time integration, the Runge-Kutta-Merson method [11] is employed. The primary variables are ρ, ρ_n, ρe, ρv_1 and ρv_2.

The rectangular computational domain Ω is covered by the orthogonal mesh depicted in Fig. 1, i.e., Ω is covered by rectangles. All of the scalar variables are defined at the vertices (referred to as s-nodes) of these rectangles. The variables ρv_1 and ρv_2 are defined at the midpoints (referred to as 1- and 2-nodes, respectively) of the vertical and horizontal edges of these rectangles, respectively. The upper indices s, 1 and 2 refer to the s-, 1- and 2-nodes. Each s-, 1- and 2-node that does not lie on $\partial\Omega$ is surrounded by a rectangular finite volume each edge of which lies on the axis of symmetry of the line segment connecting this node with a neighboring s-, 1- and 2-node, respectively (see Fig. 1).

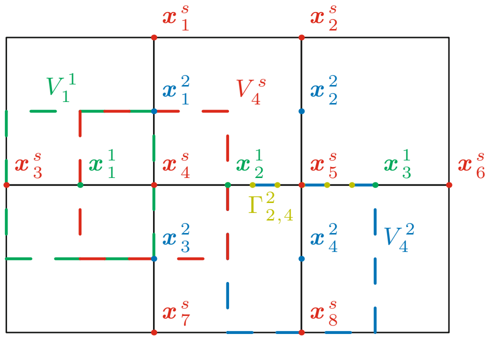


Fig. 1 Mesh of rectangles (*black lines*) and finite volumes (*dashed lines*). We have $\boldsymbol{x}^s_{4,3} = \boldsymbol{x}^1_1, \boldsymbol{x}^s_{4,1} = \boldsymbol{x}^2_1$, $\boldsymbol{x}^s_{4,5} = \boldsymbol{x}^1_2, \boldsymbol{x}^s_{4,7} = \boldsymbol{x}^2_3$ ($\Lambda^s_4 = \{1, 3, 5, 7\}$) and so on

Moreover, the mesh covering the domain Ω is extended by one layer of auxiliary rectangles [1] (the extension of Ω is denoted by $\tilde{\Omega}$), and the boundary conditions are prescribed at the corresponding s-, 1- and 2-nodes (called 'dummy nodes') lying in these rectangles. Therefore, equations (1), (2), (3), and (4) are solved in the whole of Ω.

We shall use the following notation (see Fig. 1):

- $\mathscr{X}^s = \{\boldsymbol{x}^s_i\}_{i=1}^{N_s}$, $\mathscr{X}^1 = \{\boldsymbol{x}^1_i\}_{i=1}^{N_1}$ and $\mathscr{X}^2 = \{\boldsymbol{x}^2_i\}_{i=1}^{N_2}$ and $\tilde{\mathscr{X}}^s = \{\boldsymbol{x}^s_i\}_{i=1}^{\tilde{N}_s}$, $\tilde{\mathscr{X}}^1 = \{\boldsymbol{x}^1_i\}_{i=1}^{\tilde{N}_1}$ and $\tilde{\mathscr{X}}^2 = \{\boldsymbol{x}^2_i\}_{i=1}^{\tilde{N}_2}$ are the sets of all s-, 1- and 2-nodes in Ω and in $\tilde{\Omega}$, respectively.
- $\Lambda^\alpha_i = \left\{ j \,|\, \boldsymbol{x}^\alpha_j \text{ is a neighbour of } \boldsymbol{x}^\alpha_i \right\}$.
- V^α_i is the finite volume associated with the node $\boldsymbol{x}^\alpha_i$.
- $\boldsymbol{x}^\alpha_{i,j}$ is the midpoint of the line segment connecting the nodes $\boldsymbol{x}^\alpha_i$ and $\boldsymbol{x}^\alpha_j$.
- $\Gamma^\alpha_{i,j}$ is the common face of the volumes V^α_i and V^α_j.
- $f(\boldsymbol{x}^\alpha_i) = f^\alpha_i$ and $f(\boldsymbol{x}^\alpha_{i,j}) = f^\alpha_{i,j}$, where the time coordinate is omitted.
- $[\boldsymbol{f}]_k$ denotes the k-th component of the vector $\boldsymbol{f}$, when there are too many symbols in the definition of $\boldsymbol{f}$.

The preceding notation will be used for scalar- (f) as well as vector-valued ($\boldsymbol{f}$) functions.

The numerical scheme is derived by integrating each of equations (1), (2), (3), and (4) over a corresponding volume V^α_i, applying the Green formula and using the following approximation formulas:

- $\int_{V^\alpha_i} f(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \doteq |V^\alpha_i| f^\alpha_i$, where $|V^\alpha_i|$ denotes the area of V^α_i.
- $\int_{V^\alpha_i} (\nabla f)\,(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \doteq |V^\alpha_i|\,(\nabla f)^\alpha_i$.
- $\int_{\partial V^\alpha_i} \boldsymbol{f}(\boldsymbol{x}) \cdot \boldsymbol{n}\, \mathrm{d}\boldsymbol{x} \doteq \sum_{j \in \Lambda^\alpha_i} |\Gamma^\alpha_{i,j}| \boldsymbol{f}^\alpha_{i,j} \cdot \boldsymbol{n}^\alpha_{i,j}$, where $|\Gamma^\alpha_{i,j}|$ denotes the length of $\Gamma^\alpha_{i,j}$, and $\boldsymbol{n}^\alpha_{i,j}$ is the unit outward normal with respect to $\Gamma^\alpha_{i,j}$.

We get the system of ordinary differential equations (the dummy nodes are used),

$$|V_i^s|\dot{\rho}_i^s + \textstyle\sum_{j\in\Lambda_i^\alpha} |\Gamma_{i,j}^s|\underline{\rho_{i,j}^s}\boldsymbol{v}_{i,j}^s \cdot \boldsymbol{n}_{i,j}^s = 0,\ i = 1,2,\dots N_s; \tag{12}$$

$$|V_i^s|\dot{\rho}_{n,i}^s + \textstyle\sum_{j\in\Lambda_i^\alpha} |\Gamma_{i,j}^s|\underline{\rho_{n,i,j}^s}\left(\boldsymbol{v}_{i,j}^s + \boldsymbol{V}_{n,i,j}^s\right) \cdot \boldsymbol{n}_{i,j}^s = 0,\ i = 1,2,\dots N_s; \tag{13}$$

$$\begin{gathered}|V_i^\alpha|(\dot{\rho v_\alpha})_i^\alpha + |V_i^\alpha|\left[(\nabla p)_i^\alpha\right]_\alpha + \textstyle\sum_{j\in\Lambda_i^\alpha} |\Gamma_{i,j}^\alpha|\left[(\boldsymbol{P} - p\boldsymbol{I})_{i,j}^\alpha \cdot \boldsymbol{n}_{i,j}^\alpha\right]_\alpha \\ + \textstyle\sum_{j\in\Lambda_i^\alpha} |\Gamma_{i,j}^\alpha|\underline{(\rho v_\alpha)_{i,j}^\alpha}\boldsymbol{v}_{i,j}^\alpha \cdot \boldsymbol{n}_{i,j}^\alpha = |V_i^\alpha|\rho_i^\alpha g_\alpha,\ i = 1,2,\dots N_\alpha, \alpha = 1,2;\end{gathered} \tag{14}$$

$$\begin{aligned}|V_i^s|(\dot{\rho e})_i^s + \textstyle\sum_{j\in\Lambda_i^\alpha} |\Gamma_{i,j}^s|\left[(\boldsymbol{P}\cdot\boldsymbol{v})_{i,j}^s + \boldsymbol{Q}_{i,j}^s + \underline{(\rho e)_{i,j}^s}\boldsymbol{v}_{i,j}^s\right] \cdot \boldsymbol{n}_{i,j}^s \\ = |V_i^s|\boldsymbol{g}\cdot(\rho\boldsymbol{v})_i^s,\ i = 1,2,\dots N_s;\end{aligned} \tag{15}$$

where

$$\boldsymbol{V}_{n,i,j}^s = -\textstyle\sum_{l\in\{n,g\}} D_{n,l,i,j}^s\left(\boldsymbol{d}_{l,i,j}^s + k_{T_l,i,j}^s\left(\nabla T_{i,j}^s/T_{i,j}^s\right)\right), \tag{16}$$

$$\boldsymbol{d}_{l,i,j}^s = \nabla\left(p_l/p\right)_{i,j}^s + \left(p_{l,i,j}^s/p_{i,j}^s - X_{l,i,j}^s\right)\left(\nabla p_{i,j}^s/p_{i,j}^s\right), \tag{17}$$

$$\boldsymbol{P}_{i,j}^\alpha = (p\boldsymbol{I} - 2\mu\boldsymbol{S})_{i,j}^\alpha,\ \alpha = s,1,2, \tag{18}$$

$$\boldsymbol{Q}_{i,j}^s = -\lambda\left(\nabla T\right)_{i,j}^s + p_{i,j}^s\textstyle\sum_{l\in\{n,g\}}\left(k_{T_l,i,j}^s + \frac{\kappa}{\kappa-1}\left(p_{l,i,j}^s/p_{i,j}^s\right)\right)\boldsymbol{V}_{l,i,j}^s. \tag{19}$$

For stability reasons, the term $\int_{V_i^\alpha}\nabla p$ is approximated as a volume integral in (14), and the underlined terms are modified by the full upwind formula $f_{i,j}^\alpha = f_i^\alpha$ if $\boldsymbol{v}_{i,j}^\alpha \cdot \boldsymbol{n}_{i,j}^\alpha \geq 0$; otherwise, $f_{i,j}^\alpha = f_j^\alpha$, $\alpha = s,1,2$.

In equations (12), (13), (14), and (15), only the function values that are really needed are calculated. This is realized in the following way: First, we calculate all of the values $\rho(\boldsymbol{x}_i^s)$, $i = 1,2,\dots\tilde{N}_s$. Second, we calculate all of the values $v_\alpha(\boldsymbol{x}_i^\alpha)$, $(\rho v_\alpha)(\boldsymbol{x}_i^\alpha)$, $i = 1,2,\dots\tilde{N}_\alpha$, $\alpha = 1,2$. The missing values of ρ are obtained via interpolation, where the same interpolation as in the next paragraph is employed. Third, we calculate the rest of the scalar variables at all of the s-nodes in $\tilde{\Omega}$.

In general, the function values at the midpoints of the finite volume faces are calculated via linear interpolation from the nearest function values. If the functions ρ, p and ρe are supposed to be approximately exponential in the x_i direction (e.g., due to initial conditions), their values are interpolated exponentially in this direction. The same applies for extrapolation. The fractions $\rho_k/\rho = X_n$ and p_k/p are interpolated only linearly. ρ_n is calculated from ρ and X_n. Thus, we have, for example (see Fig. 1 for the notation), $\rho_2^1 \doteq \left(\rho_4^s + \rho_5^s\right)/2$, $v_2(\boldsymbol{x}_2^1) \doteq \frac{v_{2,u}-v_{2,d}}{[\boldsymbol{x}_1^2]_2-[\boldsymbol{x}_3^2]_2}\left(\left[\boldsymbol{x}_4^s\right]_2 - \left[\boldsymbol{x}_3^2\right]_2\right) + v_{2,d}$, where $v_{2,u} = \left(v_2(\boldsymbol{x}_1^2) + v_2(\boldsymbol{x}_2^2)\right)/2$ and $v_{2,d} = \left(v_2(\boldsymbol{x}_3^2) + v_2(\boldsymbol{x}_4^2)\right)/2$.

The spatial derivatives are calculated from the nearest values as well. We have, for example, $(\partial v_1/\partial x_1)(\boldsymbol{x}_2^1) \doteq \left(v_1(\boldsymbol{x}_1^1) - v_1(\boldsymbol{x}_3^1)\right) / \left(\left[\boldsymbol{x}_1^1\right]_1 - \left[\boldsymbol{x}_3^1\right]_1\right)$.

Our experience indicates that the exponential extrapolation of ρ and p on the boundary together with the approximation of $\int_{V_i^2} [\nabla p]_2$ as a volume integral in (14) is necessary if we model a system in which the density and pressure distribution is exponential. If, for example, the linear extrapolation is employed in such a case, the numerical scheme produces non-physical oscillations of the state variables which grow without limits.

4 Numerical Test

In this section, we present results of one of our numerical tests. Describing the boundary conditions, we shall use the abbreviations *lef*, *rig*, *top*, *bot* which refer to the left, right, top and bottom edge of $\tilde{\Omega}$. The values of the physical constants used in this section are listed in Table 1.

We consider the domain $\Omega = (0.0, 3.0) \times (-0.5, 0.5)$, where the units are [m], and there are 60 and 20 squares in the x_1- and x_2-direction, respectively. The same squares are used as the auxiliary rectangles. The following hydrostatic initial ($t_{\text{ini}} = 0.0\text{s}$) conditions are considered: $\boldsymbol{v}_{\text{ini}}(\boldsymbol{x}) = \boldsymbol{v}_{\text{ref}}$, $\rho_{n,\text{ini}}(\boldsymbol{x}) = 0$, $T_{\text{ini}}(\boldsymbol{x}) = T_{\text{ref}}$, $\rho_{\text{ini}}(\boldsymbol{x}) = p_{\text{ref}} \frac{M_g}{RT_{\text{ref}}} \exp\left(\frac{M_g g_2}{RT_{\text{ref}}} x_2\right)$.

At the dummy nodes, the following setup for $\boldsymbol{v}$, ρ, ρ_n and p is employed:

- Left edge. $\boldsymbol{v}|_{\text{lef}}(\boldsymbol{x}) = \boldsymbol{v}_{\text{ref}}$, $\rho|_{\text{lef}}(\boldsymbol{x}) = (p_{\text{ref}} + 100.0) \frac{M_g}{RT_{\text{ref}}} \exp\left(\frac{M_g g_2}{RT_{\text{ref}}} x_2\right)$, $\rho_n|_{\text{lef}}(\boldsymbol{x}) = X_{n,\text{ref}} \rho|_{\text{lef}}(\boldsymbol{x})$; the pressure p is extrapolated constantly.
- Right edge. $\boldsymbol{v}|_{\text{rig}}(\boldsymbol{x}) = \boldsymbol{v}_{\text{ref}}$; the densities ρ and ρ_n and the pressure p are extrapolated constantly.
- Top and bottom edge. $\boldsymbol{v}|_{\text{top}}(\boldsymbol{x}) = \boldsymbol{v}_{\text{ref}}$, $\boldsymbol{v}|_{\text{bot}}(\boldsymbol{x}) = \boldsymbol{v}_{\text{ref}}$; the density ρ_n is calculated from X_n, which is extrapolated constantly. In accordance with the information at the end of Sect. 3, the density ρ and the pressure p are extrapolated exponentially.

The coefficient k_{Tn} is calculated using the formula $k_{Tn} = 0.35 X_n \overline{M} M_n^{-1}$, which is based on information in [2].

Table 1 Values of constant physical parameters

Par.	Value	Unit	Par.	Value	Unit
$D_{g,n}$	$-8.35 \cdot 10^{-5}$	$\text{m}^2 \cdot \text{s}^{-1}$	g_1	0.0	$\text{m} \cdot \text{s}^{-2}$
μ	$1.725 \cdot 10^{-5}$	$\text{kg} \cdot \text{m}^{-1} \cdot \text{s}^{-1}$	g_2	-9.81	$\text{m} \cdot \text{s}^{-2}$
λ	0.02428	$\text{kg} \cdot \text{m} \cdot \text{K}^{-1} \cdot \text{s}^{-3}$	p_{ref}	101,325	Pa
κ	1.4	–	T_{ref}	295.15	K
M_{g}	0.02896	$\text{kg} \cdot \text{mol}^{-1}$	$X_{n,\text{ref}}$	0.001	–
M_{n}	0.13139	$\text{kg} \cdot \text{mol}^{-1}$	$v_{\text{ref},1}$	1.0	$\text{m} \cdot \text{s}^{-1}$
R	8.3144621	$\text{J} \cdot \text{K}^{-1} \cdot \text{mol}^{-1}$	$v_{\text{ref},2}$	0.0	$\text{m} \cdot \text{s}^{-1}$

The numerical results are presented in Figs. 2 and 3. We can see how the wave of higher density and non-zero density of NAPL vapor moves towards the right edge (the final density distribution in the x_2-direction equals to the density prescribed on the left edge at every $x_1 \in [0.0, 3.0]$). Note that the wavefront in Fig. 3 seems to be skew because due to gravity, the density (unlike the mass fraction of NAPL vapor) varies in the x_2-direction.

Finally, Fig. 4 shows the density distribution calculated with the spatial mesh which has 240 and 80 squares in the x_1- and x_2-direction, respectively. Comparison between Figs. 3 and 4 indicates the numerical convergence.

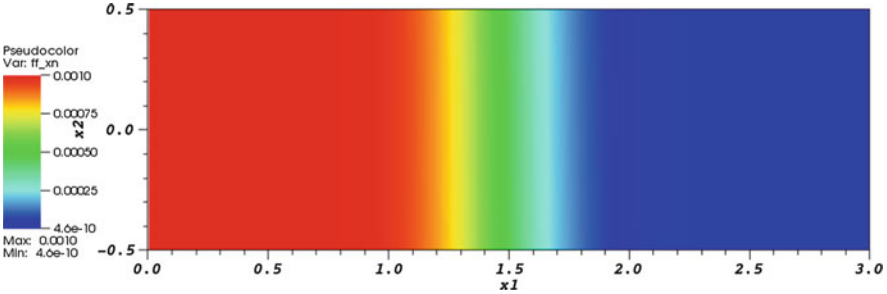


Fig. 2 Mass fraction of NAPL vapor X_n [−] on mesh 60 × 20 at time $t = 1.5$ s

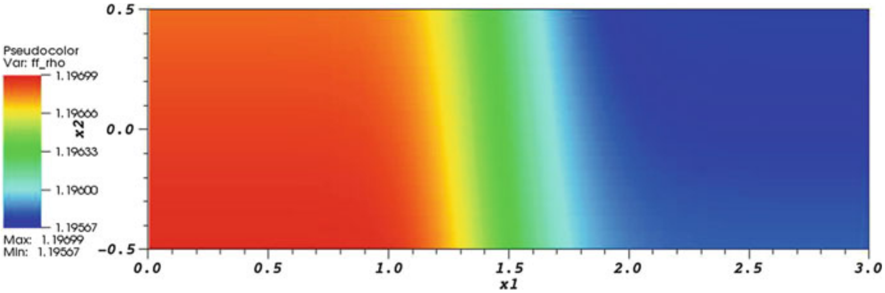


Fig. 3 Density ρ [kg · m^{-3}] on mesh 60 × 20 at time $t = 1.5$ s

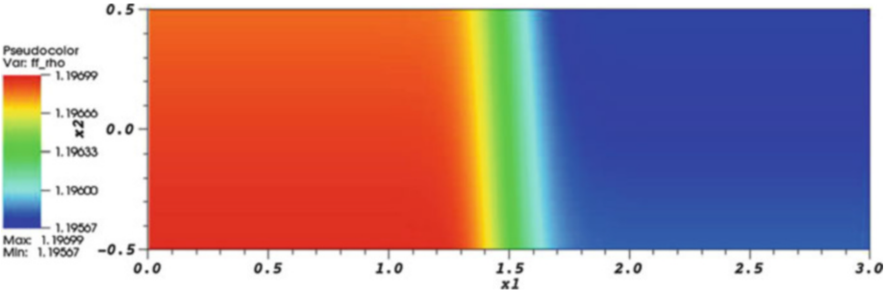


Fig. 4 Density ρ [kg · m^{-3}] on mesh 240 × 80 at time $t = 1.5$ s

5 Conclusions

In the numerical results, our numerical scheme did not produce any non-physical oscillation in the density ρ_n, mass fraction X_n and pressure p, and in the regions where the gradient of the solution is discontinuous, it adds a certain amount of artificial diffusion. Further, the modifications on the scheme mentioned at the end of the Sect. 3 seem to be necessary if we model a system in which the spatial distribution of pressure and density is exponential.

Acknowledgements The first two authors were supported by the project "Development and Validation of Porous Media Fluid Dynamics and Phase Transitions Models for Subsurface Environmental Application" of Czech Ministry of Education, Youth and Sports Kontakt II LH14003, and the third author was supported by the project "Semi-implicit methods for solution of partial differential equations" VEGA 1/0728/15.

References

1. J. Blazek, *Computational Fluid Dynamics: Principles and Applications* (Elsevier Science, Amsterdam/New York, 2001)
2. S. Chapman, T.G. Cowling, *The Mathematical Theory of Non-Uniform Gases*, 2nd edn. (Cambridge University Press, Cambridge, 1952)
3. H. Davarzani, K. Smits, R. Tolene, T. Illangasekare, Study of the effect of wind speed on evaporation from soil through integrated modeling of the atmospheric boundary layer and shallow subsurface. Water Resour. Res. **50**, 661–680 (2014)
4. J.H. Ferziger, H.G. Kaper, *Mathematical Theory of Transport Processes in Gases*, 1st edn. (North-Holland Publishing Company, Amsterdam, 1972)
5. D. Folkes, W. Wertz, J. Kurtz, T. Kuehster, Observed spatial and temporal distributions of CVOCs at Colorado and New York vapor intrusion sites. Ground Water Monit. Remediat. **29**, 70–80 (2009)
6. V. Giovangigli, *Multicomponent Flow Modeling*, 1st edn. (Birkhäuser, Boston, 1999)
7. H. Gómez, I. Colominas, F. Navarrina, M. Casteleiro, A mathematical model and a numerical model for hyperbolic mass transport in compressible flows. Heat Mass Transf. **45**, 219–226 (2008)
8. K. Mosthaf, K. Baber, B. Flemisch, R. Helmig, A. Leijnse, I. Rybak, B. Wohlmuth, A coupling concept for two-phase compositional porous-medium and single-phase compositional free flow. Water Resour. Res. **47** (2011). doi:10.1029/2011WR010685
9. O.A. Neves, E.C. Romao, J.B. Campos-Silva, Numeric simulation of pollutant dispersion by a control-volume based on finite element method. Int. J. Numer. Methods Fluids **66**, 1073–1092 (2011)
10. R. Ouzani, M. Si-Ameur, Numerical study of hydrogene–air mixing in turbulent compressible coaxial jets. Int. J. Hydrog. Energy **40**, 9539–9554 (2015)
11. O. Pártl, Computational studies of bacterial colony model. Am. J. Comput. Math. **3**, 147–157 (2013)
12. S. Patankar, *Numerical Heat Transfer and Fluid Flow* (Hemisphere Publishing Corporation, Panama, 1980)

Numerical Solution of Constrained Curvature Flow for Closed Planar Curves

Miroslav Kolář, Michal Beneš, and Daniel Ševčovič

Abstract This paper presents results of computational studies of the evolution law for the constrained mean curvature flow. The considered motion law originates in the theory of phase transitions in crystalline materials. It describes the evolution of closed embedded curves with constant enclosed area. In the paper, the motion law is treated by the parametric method, which leads into the system of degenerate parabolic equations for the parametric description of the curve. This system is numerically solved by means of the flowing finite volume method enhanced by tangential redistribution. Qualitative and quantitative results of computational studies are presented.

1 Introduction

The objective of this article is to investigate the numerical solution of non-local, area preserving curvature flow for closed planar curves. The flow is given by the following geometric evolution equation

$$v_\Gamma = -\kappa_\Gamma + F, \quad \text{where } F = \frac{1}{L(\Gamma_t)} \int_{\Gamma_t} \kappa_\Gamma \mathrm{d}s, \tag{1}$$

$$\Gamma_t|_{t=0} = \Gamma_{ini}. \tag{2}$$

Here, Γ_t is a C^1 smooth Jordan curve of the length $L(\Gamma_t) = \int_{\Gamma_t} \mathrm{d}s$ evolving in time. It is evolved in the direction of the outer normal with velocity v_Γ and driven by the curvature κ_Γ and the particular non-local force term F. Our objective is to find

M. Kolář (✉) • M. Beneš
Faculty of Nuclear Sciences and Physical Engineering, Department of Mathematics, Czech Technical University in Prague, Trojanova 13, 120 00, Prague 2, Czech Republic
e-mail: kolarmir@fjfi.cvut.cz; michal.benes@fjfi.cvut.cz

D. Ševčovič
Faculty of Mathematics, Department of Applied Mathematics and Statistics, Physics and Informatics, Comenius University, 842 48, Bratislava, Slovakia
e-mail: sevcovic@fmph.uniba.sk

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_52

a family $\{\Gamma_t : t \in (0, T_{max})\}$ of closed nonselfintersecting planar curves evolving from the initial curve Γ_{ini} according to (1). Although the evolution Eq. (1) does not involve a tangential redistribution term, any parametrization of the initial curve in (2) inherently incorporates initial redistribution of grid points which is then propagated along curve evolution.

Equation (1) belongs to a family of constrained curvature driven flows described by general evolution law

$$v_\Gamma = -\kappa_\Gamma + G,$$

where G is a possibly non-local force term preserving some quantity. In our case, the particular choice of the force term as $\int_{\Gamma_t} \kappa_\Gamma \mathrm{d}s \big/ L(\Gamma_t)$ leads to the area preserving curvature flow. Such geometric motion laws similar to (1) are discussed in the literature (see, e.g., [1–5]). Another geometric evolution laws similar to (1) treating, e.g., the length preserving curvature flow or the isoperimetric ratio gradient flow are studied and discussed in, e.g., [6].

The constrained motion driven by the curvature has also been investigated, in [7, 8] within the context of a modification of the Allen-Cahn equation (see [9, 10]). The non-local character of the geometric governing equation (1) is strongly connected with the studies of the recrystallization phenomena, where a fixed, previously melted volume of the liquid phase solidifies again (see [11]).

2 Parametric Method

The method presented in this paper is based on parametric description of the smooth time-dependent curve Γ_t ($t \geq 0$) by means of the vectorial mapping

$$\mathbf{X}(u, t) = (X_1(u, t), X_2(u, t)),$$

where $u \in [0, 1]$ is a dimensionless parameter in a given fixed interval. Throughout this paper, the parametrization is orientated counter-clockwise and periodic boundary conditions at $u = 0$ and $u = 1$ are imposed, i.e., $\mathbf{X}(0, t) = \mathbf{X}(1, t)$ and $\partial_u \mathbf{X}(0, t) = \partial_u \mathbf{X}(1, t)$.

Consequently, geometrical quantities of interest can be prescribed by the parametrization $\mathbf{X}$. The unit tangent and normal vectors $\mathbf{t}_\Gamma$ and $\mathbf{n}_\Gamma$ are defined straightforwardly, and the curvature is given by Frenet formulae:

$$\mathbf{t}_\Gamma = \frac{\partial_u \mathbf{X}}{|\partial_u \mathbf{X}|}, \quad \mathbf{n}_\Gamma = \frac{\partial_u \mathbf{X}^\perp}{|\partial_u \mathbf{X}|}, \quad \kappa_\Gamma(\mathbf{X}) = -\frac{1}{|\partial_u \mathbf{X}|} \partial_u \left(\frac{\partial_u \mathbf{X}}{|\partial_u \mathbf{X}|} \right) \cdot \mathbf{n}_\Gamma.$$

Here $\mathbf{X}^\perp = (X_2, -X_1)$. This choice is in accordance with the rule $\det(\mathbf{n}_\Gamma, \mathbf{t}_\Gamma) = 1$. Notice that in our case the curvature of the unit circle is $\kappa_\Gamma = 1$. The normal velocity is just a projection of the point velocity $\mathbf{v}_\Gamma = \partial_t \mathbf{X}$ to the normal direction $\mathbf{n}_\Gamma$, i.e., $v_\Gamma = \mathbf{v}_\Gamma \cdot \mathbf{n}_\Gamma$. Finally, the curve Γ_t evolves according to the law (1) provided the parametric mapping $\mathbf{X}$ satisfies the following system of degenerate parabolic equations

$$\partial_t \mathbf{X} = \frac{1}{|\partial_u \mathbf{X}|} \partial_u \left(\frac{\partial_u \mathbf{X}}{|\partial_u \mathbf{X}|} \right) + F \frac{\partial_u \mathbf{X}^\perp}{|\partial_u \mathbf{X}|}, \tag{3}$$

$$\mathbf{X}|_{t=0} = \mathbf{X}_{\mathrm{ini}}, \tag{4}$$

for $t \in (0, T_{max})$ and $u \in [0, 1]$. The driving force F of flow (1) written by means of the parametrization $\mathbf{X}$ becomes

$$F = \int_{\Gamma_t} \kappa_\Gamma \mathrm{d}s \Big/ L(\Gamma_t) = \int_0^1 \kappa_\Gamma(\mathbf{X}) |\partial_u \mathbf{X}| \mathrm{d}u \Big/ \int_0^1 |\partial_u \mathbf{X}| \mathrm{d}u.$$

For details on this approach, we refer he reader to, e.g., [12–15]. Another approach dealing with area preserving flows is based on the tangential velocity dependent on the Laplace-Beltrami operator acting on the curvature. For such geometric flows (see [16]) is well known that they describe area preserving geometric flows. The main advantage of this approach is in fast and straightforward numerical treatment, which is noticeable especially when comparing to other interface capturing methods, such as the level-set method [17] or the phase-field method [18]. However, this approach itself is not able to treat the cases, where changes in curve topology occurs (like merging or splitting). For such a task, separate algorithms have to be developed [19].

We denote

$$A(\Gamma_t) = \frac{1}{2} \int_0^1 \det(\mathbf{X}, \partial_u \mathbf{X}) \mathrm{d}u. \tag{5}$$

Then the flow (1) preserves the quantity $A = A(\Gamma_t)$, i.e., $A(\Gamma_t) = A(\Gamma_{ini})$ for all $t \geq 0$. For a closed curve, the quantity $A(\Gamma_t)$ represents the enclosed area. Here we remind the following result, which is known for the case when Γ_t is the Jordan curve (see e.g., [20]).

Remark 1 Let $\{\Gamma_t\}_{t \geq 0}$ be a family of C^1 smooth Jordan curves evolving in the normal direction according to (1) and parametrized by the mapping $\mathbf{X}$ satisfying (3–4). Then

$$\frac{\mathrm{d}A(\Gamma_t)}{\mathrm{d}t} = 0.$$

3 Tangential Effects

By nature of law (1), the tangential terms do not affect the shape of the curve. Hence they are not important from the analytical point of view. However, considering numerical treatment of (1), properly chosen tangential terms can significantly affect the solution. Discussion on the concept of the so called tangential redistribution can be found in, e.g., [21]. For technical details of the tangential redistribution for our parametric model, we refer the reader to [6, 12]. Notice that these papers are concerned with non-locally dependent tangential velocities. As far as locally dependent tangential velocities are concerned, we mention a tangential velocity proposed and analyzed by Dziuk and Deckelnick in [22]. Resulting parametric model (3) enhanced by tangential redistribution has the following form

$$\partial_t \mathbf{X} = \frac{1}{|\partial_u \mathbf{X}|}\partial_u\left(\frac{\partial_u \mathbf{X}}{|\partial_u \mathbf{X}|}\right) + \alpha\frac{\partial_u \mathbf{X}}{|\partial_u \mathbf{X}|} + F\frac{\partial_u \mathbf{X}^{\perp}}{|\partial_u \mathbf{X}|}, \tag{6}$$

where α is a possibly non-local function of time and curvature. In our model, we use the tangential redistribution discussed and applied in [6], which forces the discretization points to be placed asymptotically uniformly along the curve. In this case, the tangential term α satisfies

$$\frac{1}{|\partial_u \mathbf{X}|}\partial_u \alpha = \kappa_\Gamma v_\Gamma - \frac{1}{L(\Gamma_t)}\int_{\Gamma_t}\kappa_\Gamma v_\Gamma \mathrm{d}s + \omega\left(\frac{L(\Gamma_t)}{|\partial_u \mathbf{X}|} - 1\right),$$

where ω is a given scalar parameter. To ensure the uniqueness of the solution, α is required to fulfill the condition $\int_{\Gamma_t}\alpha \mathrm{d}s \big/ L(\Gamma_t) = 0$.

4 Numerical Solution

In our approach, the time evolving curve Γ_t is approximated as a piece-wise linear curve, and for the spatial discretization of governing equations (6), the flowing finite volume method is used. For technical details and discussion on the method, we refer the reader to, e.g., [5, 6, 12, 13, 21]. The discrete nodes $\mathbf{X}_i = \mathbf{X}(t, u_i)$ for $i = 0, \ldots M$ are placed along the curve Γ_t, and linear segments connecting the neighboring nodes represent the finite volumes. We denote $d_j = |\mathbf{X}_j - \mathbf{X}_{j-1}|$ for $j = 1, \ldots M$, where $\mathbf{X}_0 = \mathbf{X}_M$. Similarly to the discrete nodes $\mathbf{X}_i$, we consider discretized tangential coefficients α_i. For the way how to appropriately calculate the redistribution coefficients α_i within the context of used numerical scheme see, e.g., [12], where the problem of tangential redistribution is analyzed in detail. Finally, our semi-discrete scheme for solving (6) within the context of the motion law (1) is the following

$$\frac{d\mathbf{X}_i}{dt}\frac{d_i + d_{i+1}}{2} = \left(\frac{\mathbf{X}_{i+1} - \mathbf{X}_i}{d_{i+1}} - \frac{\mathbf{X}_i - \mathbf{X}_{i-1}}{d_i}\right) + F\frac{(\mathbf{X}_{i+1}^{\perp} - \mathbf{X}_{i-1}^{\perp})}{2}$$

$$+\alpha_i\frac{(\mathbf{X}_{i+1} - \mathbf{X}_{i-1})}{2}, \tag{7}$$

$$\kappa_i = -\frac{2}{d_i + d_{i+1}}\left(\frac{\mathbf{X}_{i+1} - \mathbf{X}_i}{d_{i+1}} - \frac{\mathbf{X}_i - \mathbf{X}_{i-1}}{d_i}\right)\cdot\frac{\mathbf{X}_{i+1}^{\perp} - \mathbf{X}_{i-1}^{\perp}}{d_i + d_{i+1}} \tag{8}$$

$$F = \frac{1}{\sum_{l=1}^{M} d_l}\sum_{l=1}^{M}\kappa_l\frac{d_{l+1} + d_l}{2}, \tag{9}$$

$$\mathbf{X}_i(0) = \mathbf{X}_{ini}(u_i), \tag{10}$$

for $i = 1, \ldots, M$. This system is solved by means of the 4th-order explicit Runge-Kutta-Merson scheme with the automatic time step (denoted as Δt_k) control and the tolerance parameter $\varepsilon = 10^{-6}$. The initial time step was chosen as h^2, where $h = 1/M$ is the mesh size dividing the parameter range [0, 1].

5 Computational Studies

We present some results of our qualitative and quantitative computational studies for the closed curves dynamics driven by (6) and treated by numerical scheme (7), (8), (9) and (10). In the following examples, we demonstrate how a solution of (6) evolves in time and approaches the circular shape.

We have measured the experimental orders of convergence (EOC) for our scheme. The measurements were performed indirectly – as the testing parameter for computation of EOC, the quantity $A(\Gamma_t)$ representing the area of the enclosed curve was chosen. We measured the differences given by the area at the initial time $A(\Gamma_{ini})$, and the areas $A(\Gamma_{T_i})$ at given data output times $T_i, i = 1, \ldots, N$. For given mesh with M segments, we evaluate the maximum and the discrete L_1 (with time steps Δt_k) norms, i.e.,

$$error_1(M) = \max_{i=1,2,\ldots N}|A(\Gamma_{ini}) - A(T_i)|,$$

$$error_2(M) = \frac{1}{T_N}\sum_{k=1}^{N}|A(\Gamma_{ini}) - A(T_i)|\Delta t_k.$$

Both errors depend on the number of finite volumes M. We estimate the order of convergence between two meshes with M_1 and M_2 volumes as

$$EOC = \log\left(error_i(M_1)/error_i(M_2)\right)\Big/\log\left(M_2/M_1\right), i = 1, 2.$$

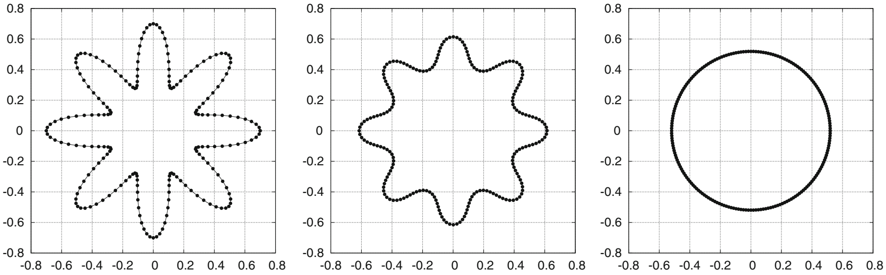


Fig. 1 The area-preserving mean curvature flow (1) in Example 1, where the initial eight-folded curve asymptotically approaches the circular shape. The curve Γ_t is depicted for time levels $t = 0$, $t = 0.005$, and $t = 0.5$

Table 1 Table of EOCs for Example 1

M	$error_1$	EOC	$error_2$	EOC
100	0.007069986241	–	0.007061259667	–
200	0.002014614725	1.8112	0.002015348016	1.8089
300	0.000944083352	1.8694	0.000945069453	1.8677
400	0.000543526916	1.9192	0.000544287109	1.9180
500	0.000352173540	1.9447	0.000352741175	1.9439

Example 1 In Fig. 1, we show the qualitative behavior of the numerical solution of problem (1), where the initial eight-folded curve Γ_{ini} is given as $\mathbf{X}(0, u) = r_{ini}(u)(\cos 2\pi u, \sin 2\pi u), u \in [0, 1]$ with r_{ini} defined as

$$r_{ini}(u) = 0.5 + 0.2\cos(16\pi u), \qquad u \in [0, 1].$$

The motion is captured in the time interval $[0, 0.5]$ and the number of finite volumes in Fig. 1 is $M = 200$. The curve Γ_t approaches the circular shape and the quantity $A(\Gamma_t)$ – the area enclosed by the curve Γ_t is preserved. The initial curve Γ_{ini} encloses the area of 0.84823 and at $t = 0.5$ the curve Γ_t encloses the area of 0.846215385275. The values of EOC for various meshes are in Table 1.

Example 2 In Fig. 2, we show the qualitative behavior of the numerical solution of problem (1), where the initial curve Γ_t with high variation of curvature is given by the parametric equations

$$\begin{aligned}\mathbf{X}(0, u) = \big(&(1 + 0.4\cos(12\pi u) + 0.2\cos(6\pi u))\cos(2\pi u),\\ &(2.5 + 0.4\sin(12\pi u) + 0.2\sin(4\pi u))\sin(2\pi u)\big) \qquad u \in [0, 1].\end{aligned}$$

The motion is captured in the time interval $[0, 5]$ and the number of finite volumes in Fig. 2 is $M = 200$. The curve Γ_t approaches the circular shape and the quantity $A(\Gamma_t)$ – the area enclosed by the curve Γ_t is preserved. The initial curve Γ_{ini} encloses

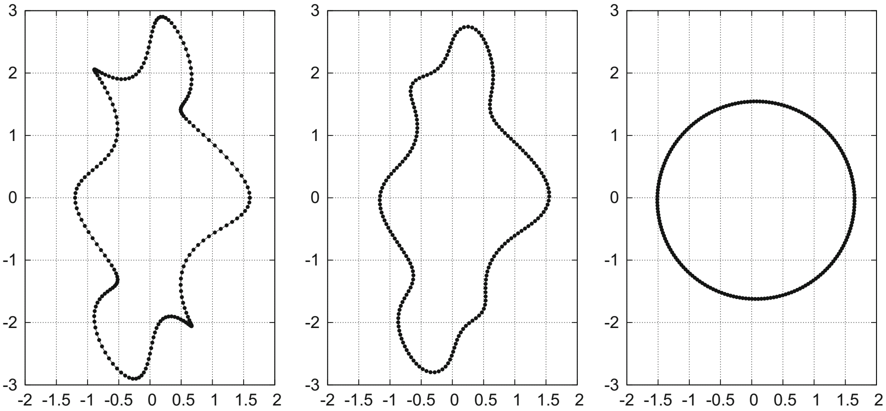


Fig. 2 The area-preserving mean curvature flow (1) in Example 2, where the initial curve asymptotically approaches the circular shape. The curve Γ_t is depicted for time levels $t = 0$, $t = 0.025$, and $t = 5$

Table 2 Table of EOCs for Example 2

M	$error_1$	EOC	$error_2$	EOC
100	0.030290732384	–	0.029986248027	–
200	0.009389794295	1.6897	0.009293578786	1.6900
300	0.004615850440	1.7514	0.004570165389	1.7505
400	0.002751185152	1.7987	0.002724720989	1.7978
500	0.001827489090	1.8333	0.001810281006	1.8324

the area of 7.85398 and at $t = 5$ the curve Γ_t encloses the area of 7.863369794295. The values of EOC for various meshes are in Table 2.

6 Conclusion

In this paper, we investigated the area-preserving mean curvature flow for closed Jordan curves in terms of qualitative and quantitative behavior of the approximate solution obtained numerically by means of the flowing finite volume method enhanced by the tangential redistribution of discretization points. Computational results suggest that the order of convergence of our numerical scheme approaches 2 in space when the convergence ratio is measured for the error measured in the enclosed area. Our studies are in agreement with theoretical indications that the solution of constrained problem (1) approaches the circular shape in steady state (see [1, 7]). This behavior corresponds to real expectations in modeling of recrystallization phenomena in solids.

Acknowledgements The first two authors were partly supported by the project No. 14-36566G *"Multidisciplinary research centre for advanced materials"* of the Grant Agency of the Czech Republic. The third author was supported by Slovak research agency grant VEGA 01/0780/15.

References

1. M. Gage, On an area-preserving evolution equation for plane curves. Contemp. Math. **51**, 51–62 (1986)
2. I.C. Dolcetta, S.F. Vita, R. March, Area preserving curve shortening flows: from phase separation to image processing. Interfaces Free Bound. **4**, 325–343 (2002)
3. C. Kublik, S. Esedoḡlu, J.A. Fessler, Algorithms for area preserving flows. SIAM J. Sci. Comput. **33**, 2382–2401 (2011)
4. J. McCoy, The surface area preserving mean curvature flow. Asian J. Math. **7**, 7–30 (2003)
5. M. Kolář, M. Beneš, D. Ševčovič, Computational studies of conserved mean-curvature flow. Mathematica Bohemica **139**(4), 677–684 (2014)
6. D. Ševčovič, S. Yazaki, On a gradient flow of plane curves minimizing the anisoperimetric ratio. IAENG Int. J. Appl. Math. **43**, 160–171 (2013)
7. J. Rubinstein, P. Sternberg, Nonlocal reaction-diffusion equations and nucleation. IMA J. Appl. Math. **48**, 249–264 (1992)
8. M. Beneš, S. Yazaki, M. Kimura, Computational studies of non-local anisotropic Allen-Cahn equation. Mathematica Bohemica **134**(4), 429–437 (2011)
9. J.W. Cahn, J.E. Hilliard, Free energy of a nonuniform system III. Nucleation of a two-component incompressible uid. J. Chem. Phys. **31**, 688–699 (1959)
10. S. Allen, J. Cahn, A microscopic theory for antiphase boundary motion and its application to antiphase domain coarsening. Acta Metallurgica **27**, 1084–1095 (1979)
11. I.V. Markov, *Crystal Growth for Beginners: Fundamentals of Nucleation, Crystal Growth, and Epitaxy*, 2nd edn. (World Scientific Publishing Company, New Jersey, 2004)
12. M. Kolář, M. Beneš, D. Ševčovič, J. Kratochvíl, Mathematical model and computational studies of discrete dislocation dynamics. IAENG Int. J. Appl. Math. **45**(3), 198–207 (2015)
13. V. Minárik, M. Beneš, J. Kratochvíl, Simulation of dynamical interaction between dislocations and dipolar loops. J. Appl. Phys. **107**, 061802 (2010)
14. K. Deckelnick, Parametric mean curvature evolution with a dirichlet boundary condition. Journal für die reine und angewandte Mathematik **459**, 37–60 (1995)
15. K. Deckelnick, G.Dziuk, C.M. Elliott, Computation of geometric partial differential equations and mean curvature flow. Acta Numerica **14**, 139–232 (2005)
16. J.W. Barrett, H. Garcke, R. Nürnberg, On the variational approximation of combined second and fourth order geometric evolution equations. SIAM J. Sci. Comput. **29**(3), 1006–1041 (2007)
17. S. Osher, J. Sethian, Fronts propagating with curvature dependent speed: algorithms based on Hamilton-Jacobi formulations. J. Comput. Phys. **79**, 12–49 (1988)
18. M. Beneš, Diffuse-interface treatment of the anisotropic mean-curvature flow. Appl. Math. **48**(6), 437–453 (2003)
19. P. Pauš, J. Kratochvíl, M. Beneš, A dislocation dynamics analysis of the critical cross-slip annihilation distance and the cyclic saturation stress in fcc single crystals at different temperatures. Acta Materialia **61**, 7917–7923 (2013)
20. D. Ševčovič, K. Mikula, Evolution of plane curves driven by a nonlinear function of curvature and anisotropy. SIAM J. Appl. Math. **61**(3), 1473–1501 (2001)
21. K. Mikula, D. Ševčovič, Computational and qualitative aspects of evolution of curves driven by curvature and external force. Comput. Vis. Sci. **6**(3), 211–225 (2004)
22. K. Deckelnick, G. Dziuk, On the approximation of the curve shortening flow. Pitman Res. Notes Math. Ser. **326**, 100–108 (1995)

Analysis of a $\boldsymbol{T}, \phi - \phi$ Formulation of the Eddy Current Problem Based on Edge Finite Elements

Alfredo Bermúdez, Marta Piñeiro, Rodolfo Rodríguez, and Pilar Salgado

Abstract The goal of this work is the analysis of a time-harmonic eddy current model with prescribed current intensities imposed on the boundary of the conducting domain. We will study a $\boldsymbol{T}, \phi - \phi$ formulation, which combines a current vector potential $\boldsymbol{T}$ with a scalar potential ϕ. A significant advantage of this method is that the expensive vector unknown $\boldsymbol{T}$ has to be computed only in conductors. Moreover, the proposed numerical method avoids the building of cutting surfaces what is very convenient in the case of complex geometries.

1 Introduction

This work deals with the solution of a time-harmonic eddy current problem defined in a three-dimensional domain; sources will be given in terms of the current intensities that cross some parts of the boundary of the conducting domain. This problem has been studied by using different unknowns (see, for instance, Chapter 8 of [1]). We will focus on the well-known $\boldsymbol{T}, \phi - \phi$ formulation, which combines a vector potential $\boldsymbol{T}$, defined only in the conducting domain, with a scalar potential ϕ, supported in the whole domain. This kind of formulation is one of the most used in commercial software for the solution of three-dimensional eddy current problems (e.g., Cedrat Flux®, ANSYS Maxwell®).

Concerning the discretization, "edge" finite elements will be employed for the approximation of the vector potential and standard Lagrange finite elements for the scalar potential. This formulation also needs the computation of a source field function in the dielectric domain, the so-called "impressed vector potential", which

A. Bermúdez • M. Piñeiro (✉) • P. Salgado
Departamento de Matemática Aplicada, Universidade de Santiago de Compostela, Santiago de Compostela, Spain
e-mail: alfredo.bermudez@usc.es; marta.pineiro@usc.es; mpilar.salgado@usc.es

R. Rodríguez
CI2MA, Departamento de Ingeniería Matemática, Universidad de Concepción, Concepción, Chile
e-mail: rodolfo@ing-mat.udec.cl

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_53

we propose to do by means of a numerical method based on the Biot-Savart law. This approach is based on the ideas proposed in [3] and eliminates the necessity of building cutting surfaces, in conductors or in dielectrics. One great advantage of this formulation is the low computational effort needed for its solution because the only vector unknown $\boldsymbol{T}$ has to be computed only in conductors, where there are generally far fewer degrees of freedom.

The outline of the paper is as follows: in Sect. 2, we derive the proposed formulation for the eddy current problem; in Sects. 3 and 4, we perform the mathematical analysis of this formulation in the continuous and discrete cases, respectively; finally, some numerical results are reported in Sect. 5.

2 $\boldsymbol{T}, \boldsymbol{\phi} - \boldsymbol{\phi}$ Formulation of the Eddy Current Problem

Eddy currents in linear, homogeneous and isotropic media are usually modeled by the low-frequency harmonic Maxwell equations,

$$\mathbf{curl}\, \boldsymbol{H} = \boldsymbol{J}, \tag{1}$$

$$i\omega\mu \boldsymbol{H} + \mathbf{curl}\, \boldsymbol{E} = \mathbf{0}, \tag{2}$$

$$\operatorname{div}(\mu \boldsymbol{H}) = 0, \tag{3}$$

along with Ohm's law

$$\boldsymbol{J} = \sigma \boldsymbol{E}, \tag{4}$$

where $\boldsymbol{E}$ is the electric field, $\boldsymbol{H}$ the magnetic field, $\boldsymbol{J}$ the current density, ω the angular frequency, μ the magnetic permeability and σ the electric conductivity. Note that the latter is non-zero only in conducting media.

Although Maxwell equations (1), (2), (3) and (4) concern the whole space, for computational purposes we restrict them to a simply connected three-dimensional bounded domain Ω, which consists of two parts, Ω_{C} and Ω_{D}, occupied by conductors and dielectrics, respectively (see Fig. 1). Domain Ω is assumed to have a Lipschitz-continuous connected boundary $\partial\Omega$ and Ω_{D} is supposed to be connected. We denote by Γ_{C}, Γ_{D} and Γ_{I} the open surfaces such that $\bar{\Gamma}_{\mathrm{C}} := \partial\Omega_{\mathrm{C}} \cap \partial\Omega$ is the outer boundary of the conductors, $\bar{\Gamma}_{\mathrm{D}} := \partial\Omega_{\mathrm{D}} \cap \partial\Omega$ that of the dielectrics and $\bar{\Gamma}_{\mathrm{I}} := \partial\Omega_{\mathrm{C}} \cap \partial\Omega_{\mathrm{D}}$ the interface between both domains. We also denote by $\boldsymbol{n}$ the outer unit normal vector to $\partial\Omega$.

The connected components of the conducting domain, Ω_{C}^{n}, $n = 1, \dots, N$, are supposed to intersect the boundary of Ω. Moreover, we assume that the outer boundary of each of them, $\partial\Omega_{\mathrm{C}}^{n} \cap \partial\Omega$, has two disjoint connected components, both of them being the closure of non-zero measure open surfaces: the "current entrances" Γ_{J}^{n}, $n = 1, \dots, N$, where the conductor is connected to an alternating

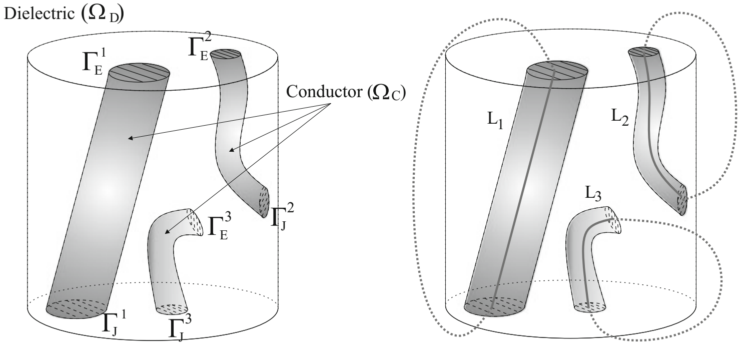


Fig. 1 Sketch of the domain (*left*). Current filaments (*right*)

electric current source, and the "current exits" $\varGamma_{\rm E}^{n}$, $n = 1, \dots, N$. We denote $\varGamma_{\rm J} := \varGamma_{\rm J}^{1} \cup \dots \cup \varGamma_{\rm J}^{N}$ and $\varGamma_{\rm E} := \varGamma_{\rm E}^{1} \cup \dots \cup \varGamma_{\rm E}^{N}$. Furthermore, we assume that $\bar{\varGamma}_{\rm J}^{n} \cap \bar{\varGamma}_{\rm J}^{m} = \emptyset$ and $\bar{\varGamma}_{\rm E}^{n} \cap \bar{\varGamma}_{\rm E}^{m} = \emptyset$, $1 \le m, n \le N$, $m \ne n$, and $\bar{\varGamma}_{\rm J} \cap \bar{\varGamma}_{\rm E} = \emptyset$.

To solve Eqs. (1), (2), (3) and (4) in a bounded domain, it is necessary to add suitable boundary conditions. We consider the following ones:

$$\int_{\varGamma_{\rm J}^{n}} \mathbf{curl}\, \boldsymbol{H} \cdot \boldsymbol{n} = I_n, \qquad n = 1, \dots, N, \tag{5}$$

$$\boldsymbol{E} \times \boldsymbol{n} = \mathbf{0} \qquad \text{on } \varGamma_{\rm E} \cup \varGamma_{\rm J}, \tag{6}$$

$$\mu \boldsymbol{H} \cdot \boldsymbol{n} = 0 \qquad \text{on } \partial\varOmega. \tag{7}$$

Conditions (5) account for the source data: the input current intensities I_n crossing each $\varGamma_{\rm J}^{n}$, $n = 1, \dots, N$. Conditions (6) and (7) will appear as natural boundary conditions of the weak formulation of the problem. The former implies the assumption that the electric current is normal to the current entrance and exit surfaces, whereas the latter means that the magnetic field is tangential to the boundary.

Our first goal is to introduce auxiliary unknowns which will be used to solve Eqs. (1), (2), (3) and (4) with boundary conditions (5), (6) and (7). First of all, note that given a complex vector of currents, $\boldsymbol{I} = (I_n)_{n=1}^{N} \in \mathbb{C}^N$, there exists $\boldsymbol{T}_0 \in \mathrm{H}(\mathbf{curl}; \varOmega)$ such that $\displaystyle\int_{\varGamma_{\rm J}^{n}} \mathbf{curl}\, \boldsymbol{T}_0 \cdot \boldsymbol{n} = I_n$ for $n = 1, \dots, N$ and $\mathbf{curl}\, \boldsymbol{T}_0 = \mathbf{0}$ in $\varOmega_{\rm D}$. Such $\boldsymbol{T}_0$ is usually called an "impressed vector potential" and can be defined in different ways (e.g., see [4]).

On the other hand, from Eq. (1), we have that $\operatorname{div} \boldsymbol{J} = 0$ in $\varOmega_{\rm C}$ and $\boldsymbol{J} \cdot \boldsymbol{n} = 0$ on $\varGamma_{\rm I}$. Therefore, $\boldsymbol{J} - \mathbf{curl}\, \boldsymbol{T}_0$ also satisfies these equations and, moreover, $\displaystyle\int_{\varGamma_{\rm J}^{n}} (\boldsymbol{J} - \mathbf{curl}\, \boldsymbol{T}_0) \cdot \boldsymbol{n} = 0$ for $n = 1, \dots, N$. Hence, it can be proved that there exists $\boldsymbol{T} \in \mathrm{H}(\mathbf{curl}; \varOmega_{\rm C})$ such that $\boldsymbol{J} - \mathbf{curl}\, \boldsymbol{T}_0 = \mathbf{curl}\, \boldsymbol{T}$ and $\boldsymbol{T} \times \boldsymbol{n} = \mathbf{0}$ on $\varGamma_{\rm I}$. Such a $\boldsymbol{T}$ is called a "current vector potential". Let $\tilde{\boldsymbol{T}} \in \mathrm{H}(\mathbf{curl}; \varOmega)$ be the extension by zero to $\varOmega$ of

$\boldsymbol{T}$. Then, $\mathbf{curl}\,\boldsymbol{H} = \boldsymbol{J} = \mathbf{curl}\,\tilde{\boldsymbol{T}} + \mathbf{curl}\,\boldsymbol{T}_0$, so that, since Ω is simply connected, $\boldsymbol{H} = \tilde{\boldsymbol{T}} + \boldsymbol{T}_0 - \mathbf{grad}\,\phi$ for some $\phi \in \mathrm{H}^1(\Omega)/\mathbb{C}$; ϕ is usually called a "magnetic scalar potential".

Taking the previous decomposition into account, the time-harmonic eddy current problem (1), (2), (3), (4), (5), (6) and (7) can be written as follows:

$$\begin{aligned}
i\omega\mu\,(\boldsymbol{T}_0 + \boldsymbol{T} - \mathbf{grad}\,\phi) + \mathbf{curl}\left(\frac{1}{\sigma}\,\mathbf{curl}(\boldsymbol{T}_0 + \boldsymbol{T})\right) &= \mathbf{0} \quad \text{in } \Omega_{\mathrm{C}},\\
\operatorname{div}\Big(\mu(\boldsymbol{T}_0 + \tilde{\boldsymbol{T}} - \mathbf{grad}\,\phi)\Big) &= 0 \quad \text{in } \Omega,\\
\left(\frac{1}{\sigma}\,\mathbf{curl}(\boldsymbol{T}_0 + \boldsymbol{T})\right) \times \boldsymbol{n} &= \mathbf{0} \quad \text{on } \Gamma_{\mathrm{E}} \cup \Gamma_{\mathrm{J}},\\
\mu(\boldsymbol{T}_0 + \tilde{\boldsymbol{T}} - \mathbf{grad}\,\phi)\cdot\boldsymbol{n} &= 0 \quad \text{on } \partial\Omega.
\end{aligned}$$

Our next goal is to introduce a weak formulation of this problem. First, let us define the following closed subspace of $\mathrm{H}(\mathbf{curl};\Omega_{\mathrm{C}})$:

$$\mathscr{Y} := \{\boldsymbol{G} \in \mathrm{H}(\mathbf{curl};\Omega_{\mathrm{C}}) \,:\, \boldsymbol{G} \times \boldsymbol{n} = \mathbf{0} \text{ on } \Gamma_{\mathrm{I}}\}\,.$$

Then, we derive the following weak form of the so called $\boldsymbol{T},\phi - \phi$ formulation:

Problem 1 Given $\boldsymbol{T}_0 \in \mathrm{H}(\mathbf{curl};\Omega)$, find $\boldsymbol{T} \in \mathscr{Y}$ and $\phi \in \mathrm{H}^1(\Omega)/\mathbb{C}$ such that

$$\begin{aligned}
&\int_{\Omega_{\mathrm{C}}} i\omega\mu(\boldsymbol{T} - \mathbf{grad}\,\phi)\cdot\bar{\boldsymbol{G}} + \int_{\Omega_{\mathrm{C}}} \frac{1}{\sigma}\,\mathbf{curl}\,\boldsymbol{T}\cdot\mathbf{curl}\,\bar{\boldsymbol{G}}\\
&\qquad = -\int_{\Omega_{\mathrm{C}}} i\omega\mu\boldsymbol{T}_0\cdot\bar{\boldsymbol{G}} - \int_{\Omega_{\mathrm{C}}} \frac{1}{\sigma}\,\mathbf{curl}\,\boldsymbol{T}_0\cdot\mathbf{curl}\,\bar{\boldsymbol{G}} \quad \forall\,\boldsymbol{G} \in \mathscr{Y},
\end{aligned}$$

$$\begin{aligned}
&-\int_{\Omega_{\mathrm{C}}} i\omega\mu\boldsymbol{T}\cdot\mathbf{grad}\,\bar{\psi} + \int_{\Omega} i\omega\mu\,\mathbf{grad}\,\phi\cdot\mathbf{grad}\,\bar{\psi}\\
&\qquad = \int_{\Omega} i\omega\mu\boldsymbol{T}_0\cdot\mathbf{grad}\,\bar{\psi} \quad \forall\,\psi \in \mathrm{H}^1(\Omega)/\mathbb{C}.
\end{aligned}$$

We advance that Problem 1 has multiple solutions $(\boldsymbol{T},\phi)$; however, $\boldsymbol{H} := \tilde{\boldsymbol{T}} + \boldsymbol{T}_0 - \mathbf{grad}\,\phi$ is uniquely determined for all of them. In the following section, we will define a well-posed auxiliary problem whose solution will lead us to obtain a particular solution of this formulation.

3 Mathematical Analysis of the $\boldsymbol{T}, \phi - \phi$ Formulation

In order to perform the analysis of the $\boldsymbol{T}, \phi - \phi$ formulation, we will write Problem 1 in terms of an auxiliary field $\widehat{\boldsymbol{H}} := \boldsymbol{H} - \boldsymbol{T}_0$ and then will apply some results from [2].

First, let us write a weak formulation of the eddy current model in terms of the impressed vector potential $\boldsymbol{T}_0$, analogous to that introduced in [2]. To this end, we define

$$\mathscr{X} := \{\boldsymbol{G} \in \mathrm{H}(\mathbf{curl}; \Omega) \ : \ \mathbf{curl}\, \boldsymbol{G} = \mathbf{0} \text{ in } \Omega_{\mathrm{D}}\}$$

and

$$\mathscr{V} := \left\{ \boldsymbol{G} \in \mathscr{X} \ : \ \langle \mathbf{curl}\, \boldsymbol{G} \cdot \boldsymbol{n}, 1 \rangle_{\Gamma_{\mathrm{J}}^n} = 0, \ n = 1, \ldots, N \right\},$$

which is a closed linear manifold of $\mathscr{X}$. The resulting formulation is as follows:

Problem 2 Given $\boldsymbol{T}_0 \in \mathrm{H}(\mathbf{curl}; \Omega)$, find $\widehat{\boldsymbol{H}} \in \mathscr{V}$ such that

$$\begin{aligned} &\int_{\Omega} i\omega\mu \widehat{\boldsymbol{H}} \cdot \bar{\boldsymbol{G}} + \int_{\Omega_{\mathrm{C}}} \frac{1}{\sigma} \mathbf{curl}\, \widehat{\boldsymbol{H}} \cdot \mathbf{curl}\, \bar{\boldsymbol{G}} \\ &\qquad = - \int_{\Omega} i\omega\mu \boldsymbol{T}_0 \cdot \bar{\boldsymbol{G}} - \int_{\Omega_{\mathrm{C}}} \frac{1}{\sigma} \mathbf{curl}\, \boldsymbol{T}_0 \cdot \mathbf{curl}\, \bar{\boldsymbol{G}} \quad \forall\, \boldsymbol{G} \in \mathscr{V}. \end{aligned}$$

By using the techniques from [2], we can prove the following result.

Theorem 3 *Problem 2 has a unique solution.*

Remark 4 If $(\boldsymbol{T}, \phi)$ is any solution to Problem 1, then it can be proved that $\widehat{\boldsymbol{H}} = \tilde{\boldsymbol{T}} - \mathbf{grad}\, \phi$ solves Problem 2. Conversely, if $\widehat{\boldsymbol{H}}$ is the solution of Problem 2, then it can be written as $\widehat{\boldsymbol{H}} = \tilde{\boldsymbol{T}} - \mathbf{grad}\, \phi$, with $\tilde{\boldsymbol{T}}$ being the extension by zero to Ω of $\boldsymbol{T} \in \mathscr{Y}$ such that $(\boldsymbol{T}, \phi)$ is a solution to Problem 1. This decomposition of $\widehat{\boldsymbol{H}}$ is not unique, unless a gauge condition is imposed. Therefore, Problem 1 is not well-posed. However, from the computational point of view, it is more interesting to obtain one particular solution of this underdetermined problem, because the more expensive vector unknown has to be computed only in conductors. Let us finally notice that the magnetic field can be subsequently computed as $\boldsymbol{H} = \boldsymbol{T}_0 + \tilde{\boldsymbol{T}} - \mathbf{grad}\, \phi$.

4 Finite Element Discretization

In this section we will discretize Problem 1 and will proceed as in the previous section for its analysis. From now on, we assume that Ω, Ω_{C} and Ω_{D} are Lipschitz polyhedra and consider regular tetrahedral meshes $\mathscr{T}_h$ of Ω such that each element $K \in \mathscr{T}_h$ is contained either in $\bar{\Omega}_{\mathrm{C}}$ or in $\bar{\Omega}_{\mathrm{D}}$ (h stands as usual for the corresponding mesh-size). Therefore, $\mathscr{T}_h(\Omega_{\mathrm{D}}) := \left\{K \in \mathscr{T}_h : K \subset \bar{\Omega}_{\mathrm{D}}\right\}$ and $\mathscr{T}_h(\Omega_{\mathrm{C}}) := \left\{K \in \mathscr{T}_h : K \subset \bar{\Omega}_{\mathrm{C}}\right\}$ are meshes of Ω_{D} and Ω_{C}, respectively.

We employ edge finite elements to approximate the current vector potential $\boldsymbol{T}$, more precisely, lowest-order Nédélec finite elements:

$$\mathscr{N}_h(\Omega_{\mathrm{C}}) := \{\boldsymbol{G}_h \in \mathrm{H}(\mathbf{curl}; \Omega_{\mathrm{C}}) : \boldsymbol{G}_h|_K \in \mathscr{N}(K)\ \forall K \in \mathscr{T}_h(\Omega_{\mathrm{C}})\},$$

where, for each tetrahedron K,

$$\mathscr{N}(K) := \left\{\boldsymbol{G}_h \in \mathbb{P}_1^3(K) : \boldsymbol{G}_h(\boldsymbol{x}) = \mathbf{a} \times \boldsymbol{x} + \mathbf{b},\ \mathbf{a}, \mathbf{b} \in \mathbb{C}^3,\ \boldsymbol{x} \in K\right\}.$$

For the magnetic potential ϕ, we use standard finite elements:

$$\mathscr{L}_h(\Omega) := \left\{\psi_h \in \mathrm{H}^1(\Omega_{\mathrm{D}}) : \psi_h|_K \in \mathbb{P}_1(K) \quad \forall K \in \mathscr{T}_h\right\}.$$

We introduce the subspace

$$\mathscr{Y}_h := \{\boldsymbol{G}_h \in \mathscr{N}_h(\Omega_{\mathrm{C}}) : \boldsymbol{G}_h \times \boldsymbol{n} = 0 \text{ on } \Gamma_{\mathrm{I}}\} \subset \mathscr{Y}$$

and a discrete impressed vector potential $\boldsymbol{T}_0^h \in \mathscr{N}_h(\Omega)$ satisfying $\mathbf{curl}\,\boldsymbol{T}_0^h = \mathbf{0}$ in Ω_{D} and $\int_{\Gamma_{\mathrm{J}}^n} \mathbf{curl}\,\boldsymbol{T}_0^h \cdot \boldsymbol{n} = I_n$ for $n = 1, \ldots, N$. We describe in Remark 8 at the end of this section how one such $\boldsymbol{T}_0^h$ can be computed in practice.

Then, the discretization of Problem 1 reads as follows:

Problem 5 Given $\boldsymbol{T}_0^h \in \mathscr{N}_h(\Omega)$, find $\boldsymbol{T}_h \in \mathscr{Y}_h$ and $\phi_h \in \mathscr{L}_h(\Omega)/\mathbb{C}$ such that

$$\int_{\Omega_{\mathrm{C}}} i\omega\mu(\boldsymbol{T}_h - \mathbf{grad}\,\phi_h) \cdot \bar{\boldsymbol{G}}_h + \int_{\Omega_{\mathrm{C}}} \frac{1}{\sigma}\,\mathbf{curl}\,\boldsymbol{T}_h \cdot \mathbf{curl}\,\bar{\boldsymbol{G}}_h$$
$$= -\int_{\Omega_{\mathrm{C}}} i\omega\mu\boldsymbol{T}_0^h \cdot \bar{\boldsymbol{G}}_h - \int_{\Omega_{\mathrm{C}}} \frac{1}{\sigma}\,\mathbf{curl}\,\boldsymbol{T}_0^h \cdot \mathbf{curl}\,\bar{\boldsymbol{G}}_h \quad \forall\, \boldsymbol{G}_h \in \mathscr{Y}_h,$$

$$-\int_{\Omega_{\mathrm{C}}} i\omega\mu\boldsymbol{T}_h \cdot \mathbf{grad}\,\bar{\psi}_h + \int_{\Omega} i\omega\mu\,\mathbf{grad}\,\phi_h \cdot \mathbf{grad}\,\bar{\psi}_h$$
$$= \int_{\Omega} i\omega\mu\boldsymbol{T}_0^h \cdot \mathbf{grad}\,\bar{\psi}_h \quad \forall\, \psi_h \in \mathscr{L}_h(\Omega)/\mathbb{C}.$$

Consider now the following subspaces:

$$\mathscr{X}_h := \{\boldsymbol{G}_h \in \mathscr{N}_h(\Omega) \, : \, \mathbf{curl}\, \boldsymbol{G}_h = \mathbf{0} \text{ in } \Omega_{\text{D}}\} \subset \mathscr{X},$$

$$\mathscr{V}_h := \left\{ \boldsymbol{G}_h \in \mathscr{X}_h \, : \, \int_{\Gamma_{\text{J}}^n} \mathbf{curl}\, \boldsymbol{G}_h \cdot \boldsymbol{n} = 0, \, n = 1, \dots, N \right\} \subset \mathscr{V}.$$

In terms of the variable $\widehat{\boldsymbol{H}}_h := \tilde{\boldsymbol{T}}_h - \mathbf{grad}\, \phi_h$ (where, as above, $\tilde{\boldsymbol{T}}_h$ is the extension to Ω by zero of $\boldsymbol{T}_h$), Problem 2 is discretized as follows:

Problem 6 Given $\boldsymbol{T}_0^h \in \mathscr{N}_h(\Omega)$, find $\widehat{\boldsymbol{H}}_h \in \mathscr{V}_h$ such that

$$\int_{\Omega} i\omega\mu \widehat{\boldsymbol{H}}_h \cdot \bar{\boldsymbol{G}}_h + \int_{\Omega_{\text{C}}} \frac{1}{\sigma} \mathbf{curl}\, \widehat{\boldsymbol{H}}_h \cdot \mathbf{curl}\, \bar{\boldsymbol{G}}_h$$
$$= - \int_{\Omega} i\omega\mu \boldsymbol{T}_0^h \cdot \bar{\boldsymbol{G}}_h - \int_{\Omega_{\text{C}}} \frac{1}{\sigma} \mathbf{curl}\, \boldsymbol{T}_0^h \cdot \mathbf{curl}\, \bar{\boldsymbol{G}}_h \quad \forall\, \boldsymbol{G}_h \in \mathscr{V}_h.$$

Following once more [2], we can prove the following results.

Theorem 7 *Problem 6 has a unique solution.*

Remark 8 As in the continuous problem, given $\boldsymbol{T}_0^h \in \mathscr{N}_h(\Omega)$, Problem 5 has at least one solution and the solution of Problem 6 can be written as $\widehat{\boldsymbol{H}}_h = \tilde{\boldsymbol{T}}_h - \mathbf{grad}\, \phi_h$ with $(\boldsymbol{T}_h, \phi_h) \in \mathscr{Y}_h \times \mathscr{L}_h(\Omega)/\mathbb{C}$ being a solution of Problem 5. This decomposition is not unique and, therefore, Problem 5 is not well posed unless a gauge condition were imposed. There are many possible ways to overcome this drawback. For example, we have solved Problem 5 by means of an iterative solver (the generalized minimal residual method).

Theorem 9 *Let $\boldsymbol{H}_h := \tilde{\boldsymbol{T}}_h - \mathbf{grad}\, \phi_h + \boldsymbol{T}_0^h$ be defined from a solution of Problem 5. Furthermore, let us assume that the magnetic field $\boldsymbol{H}$ satisfies $\boldsymbol{H}|_{\Omega_{\text{C}}} \in \mathrm{H}^r(\mathbf{curl}, \Omega_{\text{C}})$ and $\boldsymbol{H}|_{\Omega_{\text{D}}} \in \mathrm{H}^r(\Omega_{\text{D}})^3$ with $r \in \left(\frac{1}{2}, 1\right]$. Then,*

$$\|\boldsymbol{H} - \boldsymbol{H}_h\|_{\mathrm{H}(\mathbf{curl};\Omega)} \leq Ch^r \left[\|\boldsymbol{H}\|_{\mathrm{H}^r(\Omega_{\text{D}})^3} + \|\boldsymbol{H}\|_{\mathrm{H}^r(\Omega_{\text{D}})^3} \right],$$

where C is a strictly positive constant independent of h and $\boldsymbol{H}$.

Remark 10 A possible way to compute a discrete impressed vector potential $\boldsymbol{T}_0^h$ is as follows. Let $\boldsymbol{H}_{\text{BS}}$ be the Biot-Savart field in Ω corresponding to N current filaments L_n, one for each Ω_{C}^n as shown in Fig. 1, each one carrying an intensity I_n ($n = 1, \dots, N$):

$$\boldsymbol{H}_{\text{BS}}(\boldsymbol{r}) := \frac{1}{4\pi} \sum_{n=1}^{N} \int_{L_n} I_n \, dl \times \frac{\boldsymbol{r} - \boldsymbol{r}'}{|\boldsymbol{r} - \boldsymbol{r}'|^3}.$$

Then, we can take as $\boldsymbol{T}_0^h$ the field in $\mathscr{N}_h(\Omega)$ with its degrees of freedom defined for each edge ℓ of the mesh $\mathscr{T}_h$ by

$$\int_\ell \boldsymbol{T}_0^h \cdot \boldsymbol{\tau} := \begin{cases} \int_\ell \boldsymbol{H}_{\mathrm{BS}} \cdot \boldsymbol{\tau}, & \text{if } \ell \subset \bar{\Omega}_{\mathrm{D}}, \\ 0, & \text{if } \ell \subset \Omega_{\mathrm{C}} \cup \Gamma_{\mathrm{C}}, \end{cases}$$

where $\boldsymbol{\tau}$ is a unit tangent vector to ℓ. It can be checked that $\boldsymbol{H}_{\mathrm{BS}}$ satisfies $\mathbf{curl}\,\boldsymbol{H}_{\mathrm{BS}} = \mathbf{0}$ in Ω_{D} and $\int_{\Gamma_{\mathrm{J}}^n} \mathbf{curl}\,\boldsymbol{H}_{\mathrm{BS}} \cdot \boldsymbol{n} = I_n, n = 1, \ldots, N$. Moreover, it has no singularities in the computational dielectric domain Ω_{D}, since the current filaments L_n do not intersect $\bar{\Omega}_{\mathrm{D}}$. Furthermore, there are analytic expressions allowing to compute exactly the integrals above. Finally, since the resulting field $\widehat{\boldsymbol{H}}_h := \boldsymbol{H}_h - \boldsymbol{T}_0^h$ is curl-free in Ω_{D} and $\int_{\Gamma_{\mathrm{J}}^n} \mathbf{curl}\,\widehat{\boldsymbol{H}}_h \cdot \boldsymbol{n} = 0$ for $n = 1, \ldots, N$, there is no need to include cutting surfaces in the domain, even though Ω_{D} is not simply connected.

5 Numerical Results

In this section we report the numerical results obtained for an academic test that confirm the results stated in Remark 8 and the convergence of the proposed methodology.

We take as conducting domain, a piece of an infinite cylinder with radius R as shown in Fig. 2 (left), composed by a conducting material with electric conductivity σ carrying an alternating current $I(t) = I_0 \cos(\omega t)$, surrounded by dielectric material. We can obtain the analytical solution of the associated eddy current

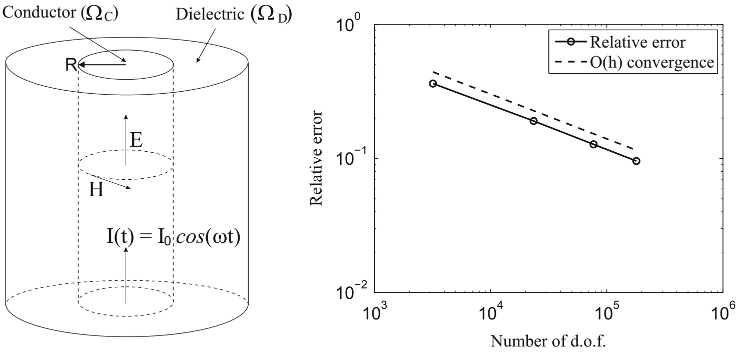


Fig. 2 Section of an infinite cylinder carrying an alternating current (*left*). Convergence order in H(**curl**; Ω) (*right*)

problem, which is:

$$\boldsymbol{H}(\boldsymbol{x}) = \begin{cases} \frac{I_0 \mathscr{I}_1(\sqrt{i\omega\mu\sigma}\rho)}{2\pi R \mathscr{I}_1(\sqrt{i\omega\mu\sigma}R)} \boldsymbol{e}_\theta, & \text{if } \rho \leq R, \\ \frac{I_0}{2\pi\rho} \boldsymbol{e}_\theta, & \text{if } \rho > R, \end{cases}$$

where $\mathscr{I}_1$ is the modified Bessel function of the first kind, and $\rho = \sqrt{x_1^2 + x_2^2}$ and $\boldsymbol{e}_\theta := (-x_2, x_1, 0)/\rho$ are the radial coordinate and the angular unit vector in cylindrical coordinates, respectively.

When comparing the numerical solution obtained from an implementation of Problem 5 with the exact one, we obtain the error curve shown in Fig. 2 (right), which shows that an order of convergence $O(h)$ is clearly attained in this case, in agreement with the theoretical results. Note that an arbitrary dashed line whose slope corresponds to the theoretical order of convergence $O(h)$ is included to allow for comparison.

References

1. A. Alonso-Rodríguez, A. Valli, *Eddy Current Approximation of Maxwell Equations: Theory, Algorithms and Applications* (Springer, Milan, 2010)
2. A. Bermúdez, R. Rodríguez, P. Salgado, Numerical solution of Eddy current problems in bounded domains using realistic boundary conditions. Comput. Methods Appl. Mech. Eng. **194**(2), 411–426 (2005)
3. O. Bíró, K. Preis, Generating source field functions with limited support for edge finite-element eddy current analysis. IEEE Trans. Mag. **43**(4), 1165–1168 (2007)
4. O. Bíró, P. Böhm, K. Preis, G. Wachutka, Edge finite element analysis of transient skin effect problems. IEEE Trans. Mag. **36**(4), 835–838 (2000)

Two Variants of Stabilized Nodal-Based FEM for the Magnetic Induction Problem

Utku Kaya, Benjamin Wacker, and Gert Lube

Abstract We consider the time-dependent magnetic induction model as a step towards the resistive magnetohydrodynamics (MHD) model in incompressible media. Conforming nodal-based finite element (FE) approximations of the induction model with Taylor-Hood type FE as well as equal-order FE for the magnetic field and the magnetic pseudo-pressure are investigated. We consider a stabilized nodal-based FEM for the numerical solution. Error estimates are given for the semidiscrete model in space. Finally, we present results for the magnetic flux expulsion problem.

1 Introduction

We consider the numerical approximation of the induction equation

$$\varrho\partial_t\mathbf{b} + \lambda\nabla\times\nabla\times\mathbf{b} + \nabla r - \nabla\times(\mathbf{u}\times(\varrho\mathbf{b})) = \mathbf{f_b}, \qquad \nabla\cdot\mathbf{b} = 0 \tag{1}$$

for the magnetic field $\mathbf{b}$ and the magnetic pseudo-pressure r with given flow field $\mathbf{u}$, force term $\mathbf{f_b}$, current density ϱ and magnetic diffusivity λ. The standard approach to the numerical solution of (1) consists of curl-conforming FEM, see [10], but has disadvantages in implementation [11]. Thus nodal elements are preferable due to their efficiency and implementation convenience. The paper [5] caused a revival of nodal-based FEM for the Maxwell problem. Stabilization techniques of residual type based on nodal-based FEM were considered for the Maxwell problem, e.g. in [1] and [3]. Extensions of nodal-based stabilized FEM to the resistive MHD model can be found in [2] where the focus is on equal-order interpolation of all unknowns.

U. Kaya (✉)
Mathematisches Seminar, Christian-Albrechts University of Kiel, Westring 383, D-24118 Kiel, Germany
e-mail: kaya@math.uni-kiel.de

B. Wacker • G. Lube
NAM, Georg-August University of Göttingen, Lotzestr. 16-18, D-37073 Göttingen, Germany
e-mail: b.wacker@math.uni-goettingen.de; lube@math.uni-goettingen.de

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_54

Here the goal is to provide a unique approach to nodal-based FE-methods with conforming Taylor-Hood type and equal-order element pairs for the approximation of $(\mathbf{b}, r)$ in problem (1). These discrete models are augmented by standard global stabilization techniques for the divergence-free constraint.

2 Continuous Magnetic Induction Equation

The norm of a Banach space X is denoted by $\|\cdot\|_X$. X' denotes the dual space of X. We denote by $(\cdot,\cdot)_G$ the inner product in $L^2(G)$ and by $\|\cdot\|_{W^{m,p}(G)}, m \in \mathbb{N}_0, p \geq 1$ the norm on the Sobolev space $W^{m,p}(G)$ for subdomains $G \subseteq \Omega$ of a bounded Lipschitz polyhedral domain $\Omega \subset \mathbb{R}^d, d \in \{2,3\}$. In case of $G = \Omega$ we will omit index Ω. Moreover, we use the spaces

$$\begin{aligned} H(\mathbf{curl};\Omega) :=& \{\mathbf{v} \in L^2(\Omega)^d \text{ s.t. } \nabla \times \mathbf{v} \in L^2(\Omega)^d\}, \\ H_0(\mathbf{curl};\Omega) :=& \{\mathbf{v} \in H(\mathbf{curl};\Omega) \text{ s.t. } \mathbf{n} \times \mathbf{v} = \mathbf{0} \text{ on } \partial\Omega\}, \\ H(\text{div};\Omega) :=& \{\mathbf{v} \in L^2(\Omega)^d \text{ s.t. } \nabla \cdot \mathbf{v} \in L^2(\Omega)\}, \\ H(\text{div } 0;\Omega) :=& \{\mathbf{v} \in H(\text{div};\Omega) \text{ s.t. } \nabla \cdot \mathbf{v} = 0\}. \end{aligned}$$

$L^2(a,b;X^d)$, $(a,b) \subset \mathbb{R}$ is the completion of $C([a,b];X^d)$ with $\|\mathbf{v}\|_{L^2(a,b;X^d)} = \big((\mathbf{v},\mathbf{v})_{L^2(a,b;X^d)}\big)^{\frac{1}{2}}$ induced by $(\mathbf{v},\mathbf{w})_{L^2(a,b;X^d)} = \int_a^b \big((\mathbf{v},\mathbf{w})_{X^d}\big)\ dt$. Moreover, $L^\infty(a,b;X^d)$ is induced by the norm $\|\mathbf{v}\|_{L^\infty(a,b;X^d)} = \sup_{a \leq t \leq b} \|\mathbf{v}(t)\|_{X^d}$.

Consider problem (1) as saddle-point problem for the magnetic field $\mathbf{b}$ and the pseudo-pressure r as Lagrange multiplier for the divergence-free constraint: Find a pair $(\mathbf{b}, r)$ satisfying

$$\varrho\partial_t\mathbf{b} - \nabla \times (\mathbf{u} \times \varrho\mathbf{b}) + \lambda\nabla \times (\nabla \times \mathbf{b}) + \nabla r \ = \ \mathbf{f}, \qquad \nabla \cdot \mathbf{b} \ = \ 0. \tag{2}$$

The initial condition $\mathbf{b}(\mathbf{x},0) = \mathbf{b}_0(\mathbf{x})$ is required to be solenoidal. For simplicity, we use the solution spaces $\mathbf{C} := H_0(\mathbf{curl};\Omega)$ and $S := H_0^1(\Omega)$ supplemented with the norms

$$\begin{aligned} \|\mathbf{c}\|_{\mathbf{C}} \ :=& \ \sqrt{\lambda}\Big(\ell^{-1}\|\mathbf{c}\|_{L^2(\Omega)^d} + \|\nabla \times \mathbf{c}\|_{L^2(\Omega)^d}\Big), \\ \|s\|_S \ :=& \ \lambda^{-\frac{1}{2}}\Big(\|s\|_{L^2(\Omega)} + \ell\|\nabla s\|_{L^2(\Omega)^d}\Big) \end{aligned}$$

where $\ell = \ell(\Omega)$ ensures dimensional consistency of the norms.

In problem (2) we multiply with test functions $\mathbf{c} \in \mathbf{C}$ resp. $q \in S$, integrate over Ω and several terms by parts, and impose the given boundary conditions. The curl-gradient formulation reads: Find $\mathbf{b} : [0,T] \to \mathbf{C},\ r : [0,T] \to S$ s.t.

$$(\varrho\partial_t\mathbf{b},\mathbf{c}) - (\mathbf{u}\times\varrho\mathbf{b},\nabla\times\mathbf{c}) + (\lambda\nabla\times\mathbf{b},\nabla\times\mathbf{c}) + (\nabla r,\mathbf{c}) \;=\; (\mathbf{f},\mathbf{c}), \tag{3a}$$

$$-\,(\nabla s,\mathbf{b}) \;=\; 0 \tag{3b}$$

for all $(\mathbf{c},s) \in (\mathbf{C},S)$ a.e. in $(0,T)$ with the assumption $\mathbf{f} \in H(\mathrm{div}\ 0;\Omega)$. We introduce the bilinear forms $a : \mathbf{C}\times\mathbf{C} \to \mathbb{R}$ and $b : \mathbf{C}\times S \to \mathbb{R}$ as

$$a(\mathbf{b},\mathbf{c}) \;:=\; (\lambda\nabla\times\mathbf{b},\nabla\times\mathbf{c}) - (\mathbf{u}\times\varrho\mathbf{b},\nabla\times\mathbf{c}), \qquad b(\mathbf{c},r) \;:=\; (\nabla r,\mathbf{c})$$

and $c(\mathbf{b},r;\mathbf{c},s) := a(\mathbf{b},\mathbf{c}) + b(\mathbf{c},r) - b(\mathbf{b},s)$. Then the weak form of (3) reads:

$$\varrho(\partial_t\mathbf{b},\mathbf{c}) + a(\mathbf{b},\mathbf{c}) + b(\mathbf{c},r) \;=\; (\mathbf{f},\mathbf{c}) \qquad \forall\mathbf{c}\in\mathbf{C}, \tag{4a}$$

$$-\,b(\mathbf{b},s) \;=\; 0 \qquad \forall s\in S, \tag{4b}$$

and in compact form as $\rho(\partial_t\mathbf{b},\mathbf{c}) + c(\mathbf{b},r;\mathbf{c},s) = (\mathbf{f},\mathbf{c}), \quad \forall(\mathbf{c},s)\in\mathbf{C}\times S$.

The choice of function spaces above yields $r \equiv 0$ a.e. in Ω. The well-posedness of (4) follows from Lemma 3 and Theorem 4 in [8].

Lemma 1 *Let* $\mathbf{u}\in[L^\infty(\Omega)]^d$. *The bilinear form a satisfies Gårding's inequality*

$$a(\mathbf{b},\mathbf{b}) \geq \gamma\|\mathbf{b}\|_{\mathbf{C}}^2 - \kappa\|\mathbf{b}\|_{L^2(\Omega)^d}^2 \qquad \text{for all } \mathbf{b}\in\mathbf{C} \tag{5}$$

with constants $\gamma := \frac{1}{4}$ *and* $\kappa := \frac{\lambda}{2\ell^2}\Big(1 + \Big[\frac{\varrho\|\boldsymbol{u}\|_{L^\infty(\Omega)^d}\ell}{\lambda}\Big]^2\Big)$. *Furthermore, it is bounded, i.e.* $|a(\mathbf{b},\mathbf{c})| \leq M\|\mathbf{b}\|_{\mathbf{C}}\|\mathbf{c}\|_{\mathbf{C}}$ *for all* $\mathbf{b},\mathbf{c}\in\mathbf{C}$.

Theorem 2 *For problem* (3) *with* $\mathbf{f}\in L^2(0,T;L^2(\Omega)^d)$ *and* $\mathbf{b}(0)\in L^2(\Omega)^d$, *we obtain the estimate*

$$\varrho\|\mathbf{b}(t)\|_{L^2(\Omega)^d}^2 + 2\gamma\int_0^t e^{\frac{3\kappa}{\varrho}(t-\tau)}\|\mathbf{b}(\tau)\|_{\mathbf{C}}^2 d\tau \leq \varrho e^{\frac{3\kappa}{\varrho}t}\|\boldsymbol{b}(0)\|_{L^2(\Omega)^d}^2 + \frac{1}{\kappa}\int_0^t e^{\frac{3\kappa}{\varrho}(t-\tau)}\|\mathbf{f}(\tau)\|_{L^2(\Omega)^d}^2 d\tau.$$

Let $\mathbf{V} = \mathbf{C}\cap H(\mathrm{div}\,0;\Omega)$ and $\mathbf{H} = H(\mathrm{div}\,0;\Omega)$. Since $b(\mathbf{c},s) = -(\nabla s,\mathbf{c}) = (s,\nabla\cdot\mathbf{c}) = 0$ for all $\mathbf{c}\in\mathbf{V}$, problem (3) is equivalent to the problem with built-in constraint: For given $\mathbf{b}_0\in\mathbf{H}$, $\mathbf{f}\in L^2(0,T;\mathbf{V}')$, find $\mathbf{b} : [0,T]\to\mathbf{V}$ s.t.

$$(\varrho\partial_t\mathbf{b},\mathbf{c}) + a(\mathbf{b},\mathbf{c}) \;=\; (\mathbf{f},\mathbf{c}) \qquad \forall\mathbf{c}\in\mathbf{V}, \ \text{ a.e. in } [0,T]. \tag{6}$$

Moreover, $\mathbf{V} \subseteq \mathbf{H} \equiv \mathbf{H}' \subseteq \mathbf{V}'$ form a Gelfand-triple: $\{\mathbf{V}, \|\cdot\|_{\mathbf{C}}\}$ and $\{\mathbf{H}, \|\cdot\|_{L^2(\Omega)^d}\}$ are Hilbert spaces. Corollary 3.49 in [10] states

$$\|\mathbf{b}\|_{L^2(\Omega)^d} \leq C_F \|\mathbf{b}\|_{\mathbf{C}} \qquad \forall \mathbf{b} \in \mathbf{V}. \tag{7}$$

Thus, $\mathbf{V}$ is continuously embedded in $\mathbf{H}$. Now, Lemma 1 allows the application of the main theorem for linear evolution problems, see [12], Theorem 23.A. This implies that for (6) there exists a unique solution $\mathbf{u} \in \mathscr{W}(0,T;\mathbf{V}) := \{\mathbf{u} \in L^2(0,T;\mathbf{V}) : \exists \mathbf{u}' \in L^2(0,T;\mathbf{V}')\}$. Hence, we obtain an existence result for the curl-gradient formulation (3), see Theorem 8 of [8].

Theorem 3 *For* $\mathbf{f} \in L^2(0,T;L^2(\Omega)^d)$ *and* $\mathbf{b}_0 = \mathbf{b}(0) \in H(\text{div } 0;\Omega)$, *there exists a unique solution* $(\mathbf{b},\ r) \in \mathscr{W}(0,T;\mathbf{C}) \times L^2(0,T;S)$ *of problem* (3).

3 Discretization of the Induction Problem

Let $\mathscr{T}_h = \{K_i\}_{i=1}^M$ be a non-overlapping admissible decomposition of the bounded polyhedron $\overline{\Omega}$ into convex polyhedral subdomains K_i s.t. $\overline{\Omega} = \cup_{j=1}^M \overline{K}_j$ with elements K_i of diameter $h_i = \text{diam}(K_i)$ and $h = \max_{i=1,\ldots,M} h_i$. Let $\mathscr{T}_h$ be shape-regular. We consider nodal-based FE-spaces with

$$\mathscr{N}_k(\Omega) = \{v_h \in \mathscr{C}^0(\overline{\Omega}) \text{ s.t. } v_h|_K \in \mathscr{P}_k(K) \quad \forall K \in \mathscr{T}_h\}. \tag{8}$$

with the set $\mathscr{P}_k(K)$ of polynomials of maximal degree $k \in \mathbb{N}$.

Remark 4 Please note that the following numerical analysis can be similarly performed for quadrilateral and hexahedral elements.

For the discrete magnetic field and pseudo-pressure, we apply Taylor-Hood type pairs $(\mathbf{b}_h, r_h) \in \mathbf{C}_h = \big([\mathscr{N}_{k+1}(\Omega)]^d \cap \mathbf{C}\big) \times \big([\mathscr{N}_k(\Omega)]^d \cap S\big)$ or equal-order pairs $(\mathbf{b}_h, r_h) \in \mathbf{C}_h = \big([\mathscr{N}_k(\Omega)]^d \cap \mathbf{C}\big) \times \big([\mathscr{N}_k(\Omega)]^d \cap S\big)$ with $k \geq 1$. Such H^1-conforming Galerkin ansatz requires stabilization since (i) a discrete inf-sup condition in subspaces of $C \times S$ is not known and (ii) the approximation of singular solutions $\mathbf{b} \in V \cap H(\text{div } 0;\Omega)$ with $\mathbf{b} \notin V \cap H^1(\Omega)^d$ is not possible, see [1]. A potential stabilized problem is:

Find $\mathbf{b}_h \in L^2(0,T;\mathbf{C}_h)$ with $\mathbf{b}_h' \in L^2(0,T;\mathbf{C}_h')$ and $r_h \in L^2(0,T;S_h)$ s.t.

$$(\rho\partial_t \mathbf{b}_h, \mathbf{c}_h) + a(\mathbf{b}_h, \mathbf{c}_h) + b(\mathbf{c}_h, r_h) \ + s_b(\mathbf{b}_h, \mathbf{c}_h) \ = \ (\mathbf{f}, \mathbf{c}_h) \ \forall \mathbf{c}_h \in \mathbf{C}_h, \tag{9a}$$

$$-\, b(\mathbf{b}_h, s_h) \ + s_r(r_h, s_h) \ = \ 0 \quad \forall s_h \in S_h, \tag{9b}$$

a.e. in $(0, T)$ with the stabilization terms

$$s_b(\mathbf{b}_h, \mathbf{c}_h) := \sum_{K \in \mathscr{T}_h} \tau_{\mathbf{b}} (\nabla \cdot \mathbf{b}_h, \nabla \cdot \mathbf{c}_h)_K, \quad s_r(r_h, s_h) := \sum_{K \in \mathscr{T}_h} \tau_r (\nabla r_h, \nabla s_h)_K.$$

We set $c_s(\mathbf{b}_h, r_h; \mathbf{c}_h, s_h) := c(\mathbf{b}_h, r_h; \mathbf{c}_h, s_h) + s_b(\mathbf{b}_h, \mathbf{c}_h) + s_r(r_h, s_h)$. In the stability analysis we apply the mesh-dependent semi-norm

$$|||(\mathbf{c}_h, s_h)|||_h := \Big(\sum_{K \in \mathscr{T}_h} \tau_{\mathbf{b}} \|\nabla \cdot \mathbf{c}_h\|_K^2 + \sum_{K \in \mathscr{T}_h} \tau_r \|\nabla s_h\|_K^2 \Big)^{\frac{1}{2}}.$$

Lemma 5 *Let* $\mathbf{f} \in L^2(0, T; L^2(\Omega)^d)$ *and* $\mathbf{b}_h(0) \in L^2(\Omega)^d$. *Then, for* $t \in [0, T]$ *and with* $C_u := 3\kappa/\varrho$, *we have*

$$\varrho \|\mathbf{b}_h(t)\|_{L^2(\Omega)^d}^2 + \int_0^t 2e^{C_u(t-\tau)} \Big(\|\mathbf{b}_h(\tau)\|_{\mathbf{C}}^2 + |||(\mathbf{b}_h(\tau), r_h(\tau))|||_h^2 \Big)\, d\tau$$
$$\le \varrho e^{C_u t} \|\mathbf{b}_h(0)\|_{L^2(\Omega)^d}^2 + \int_0^t \kappa^{-1} e^{C_u(t-\tau)} \|\mathbf{f}(s)\|_{L^2(\Omega)^d}^2 \, d\tau.$$

Proof One can proceed in $\mathbf{C}_h \times S_h$ as in the proof of Theorem 2 with additional control of the stabilization term $|||\cdot|||_h$ on the L.H.S. □

Lemma 6 *There exists a unique solution* $(\mathbf{b}_h, r_h)$ *of problem* (9).

Proof Equation (9b) with $s_h = r_h$ yields $-(\mathbf{b}_h, \nabla r_h) + \sum_K \tau_r \|\nabla r_h\|_{L^2(K)}^2 = 0$. Assuming $\min_K \tau_r \ge \tau_0 > 0$ and using the definition of norm $\|\cdot\|_{\mathbf{C}}$, we obtain

$$\|\nabla r_h\|_{L^2(\Omega)} \le \tau_0^{-\frac{1}{2}} \|\mathbf{b}_h\|_{L^2(\Omega)} \le \ell(\lambda \tau_0)^{-\frac{1}{2}} \|\mathbf{b}_h\|_{\mathbf{C}}. \tag{10}$$

Hence there exists an invertible $G_h : \mathbf{C}_h \to S_h$ with $r_h = G_h \mathbf{b}_h$. Now we add (9a) and (9b) with $(\mathbf{c}_h, s_h) = (\mathbf{b}_h, G_h \mathbf{b}_h)$. Then Lemma 5 implies

$$\varrho \|\mathbf{b}_h(t)\|_{L^2(\Omega)^d}^2 + \int_0^t 2e^{C_u(t-\tau)} \Big(\|\mathbf{b}_h(\tau)\|_{\mathbf{C}}^2 + |||(\mathbf{b}_h(\tau), G_h \mathbf{b}_h(\tau))|||_h^2 \Big)\, d\tau$$
$$\le \varrho e^{C_u t} \|\mathbf{b}_h(0)\|_{L^2(\Omega)^d}^2 + \int_0^t \kappa^{-1} e^{C_u(t-\tau)} \|\mathbf{f}(s)\|_{L^2(\Omega)^d}^2 \, d\tau. \tag{11}$$

The Cauchy-Lipschitz theorem yields existence and uniqueness of $\mathbf{b}_h : [0, T] \to V_h$. Finally, (10) guarantees existence and uniqueness of $r_h : [0, T] \to S_h$. □

Note that full control of ∇r_h is essential in order to enforce the condition $\nabla \cdot \mathbf{b} = 0$. From (10), we come up with the suggestion $\tau_r \sim \ell^2 \lambda^{-1}$.

Let $(\mathbf{b}, r)$ be the solution of (3) and $(\mathbf{b}_h, r_h)$ be the solution of (9). Then, by subtracting (9) from (3), we obtain Galerkin orthogonality

$$\varrho(\partial_t(\mathbf{b}-\mathbf{b}_h), \mathbf{c}_h) + c_s(\mathbf{b}-\mathbf{b}_h, r-r_h; \mathbf{c}_h, s_h) = s_b(\mathbf{b}, \mathbf{c}_h) + s_p(r, s_h) = 0 \tag{12}$$

for any $(\mathbf{c}_h, r_h) \in \mathbf{C}_h \times S_h$ since $r \equiv 0$ and $\nabla \cdot \mathbf{b} = 0$. Let $\mathbf{j}^{\mathbf{b}}$ and j^r be appropriate interpolation operators in $\mathbf{C}_h \times S_h$. We decompose the errors as

$$\begin{aligned} \mathbf{b}-\mathbf{b}_h &= (\mathbf{b}-\mathbf{j}^{\mathbf{b}}\mathbf{b}) + (\mathbf{j}^{\mathbf{b}}\mathbf{b}-\mathbf{b}_h) \equiv \varepsilon_{\mathbf{b}} + \mathbf{e}_{\mathbf{b}}, \\ r-r_h &= (r-j^r r) + (j^r r - r_h) \ \equiv \ \varepsilon_r + e_r. \end{aligned}$$

Set $(\mathbf{c}_h, r_h) = (\mathbf{e}_{\mathbf{b}}, e_r)$ in (12). Then we write

$$\frac{\varrho}{2}\frac{d}{dt}\|\mathbf{e}_{\mathbf{b}}\|^2_{L^2(\Omega)^d} + c_s(\mathbf{e}_{\mathbf{b}}, e_r; \mathbf{e}_{\mathbf{b}}, e_r) = -(\partial_t\varepsilon_{\mathbf{b}}, \mathbf{e}_{\mathbf{b}}) - I - II$$

with $I = c(\varepsilon_{\mathbf{b}}, \varepsilon_r; \mathbf{e}_{\mathbf{b}}, e_r), II = s_b(\varepsilon_{\mathbf{b}}, \mathbf{e}_{\mathbf{b}}) + s_r(\varepsilon_r, e_r)$; thus $r = j_r r = 0$ yields

$$\begin{aligned} I =& \Big[\sqrt{\lambda}\|\nabla\times\varepsilon_{\mathbf{b}}\|_{L^2(\Omega)^d} + \Big(\sum_K \varrho^2\|\mathbf{u}\|^2_{L^\infty(K)}\lambda^{-1}\|\varepsilon_{\mathbf{b}}\|^2_{L^2(K)}\Big)^{\frac{1}{2}}\Big]\sqrt{\lambda}\|\nabla\times\mathbf{e}_{\mathbf{b}}\|_{L^2(\Omega)^d} \\ &+ \Big(\sum_K \tau_r^{-1}\|\varepsilon_{\mathbf{b}}\|^2_{L^2(K)}\Big)^{\frac{1}{2}} |||(\mathbf{e}_{\mathbf{b}}, e_r)|||_h, \\ |II| =& \sum_K \tau_{\mathbf{b}}(\nabla\cdot\varepsilon_{\mathbf{b}}, \nabla\cdot\mathbf{e}_{\mathbf{b}})_K + \sum_K \tau_r(\nabla\varepsilon_r, \nabla e_r)_K \le |||(\varepsilon_{\mathbf{b}}, \varepsilon_r)|||_h |||(\mathbf{e}_{\mathbf{b}}, e_r)|||_h. \end{aligned}$$

So we obtain

$$\begin{aligned} &\frac{\varrho}{2}\frac{d}{dt}\|\mathbf{e}_{\mathbf{b}}\|^2_{L^2(\Omega)^d} + \lambda\|\nabla\times\mathbf{e}_{\mathbf{b}}\|^2_{L^2(\Omega)^d} - \varrho\|\mathbf{u}\times\mathbf{e}_{\mathbf{b}}\|_{L^2(\Omega)^d}\|\nabla\times\mathbf{e}_{\mathbf{b}}\|_{L^2(\Omega)^d} \\ &+ |||(\mathbf{e}_{\mathbf{b}}, e_r)|||^2_h \le S_1\|\mathbf{e}_{\mathbf{b}}\|_{L^2(\Omega)^d} + S_2\sqrt{\lambda}\|\nabla\times\mathbf{e}_{\mathbf{b}}\|_{L^2(\Omega)^d} + S_3|||(\mathbf{e}_{\mathbf{b}}, e_r)|||_h \end{aligned} \tag{13}$$

with $S_1 = \|\partial_t\varepsilon_{\mathbf{b}}\|_{L^2(\Omega)^d}$, $S_2 = \sqrt{\lambda}|\nabla\times\varepsilon_{\mathbf{b}}\|_{L^2(\Omega)^d} + \Big(\sum_K \frac{\varrho^2\|\mathbf{u}\|^2_{L^\infty(K)}}{\lambda}\|\varepsilon_{\mathbf{b}}\|^2_{L^2(K)}\Big)^{\frac{1}{2}}$, $S_3 = |||(\varepsilon_{\mathbf{b}}, \varepsilon_r)|||_h + \Big(\sum_K \tau_r^{-1}\|\varepsilon_{\mathbf{b}}\|^2_{L^2(K)}\Big)^{\frac{1}{2}}$. Young's inequality in (13) gives

$$\begin{aligned} &\varrho\frac{d}{dt}\|\mathbf{e}_{\mathbf{b}}\|^2_{L^2(\Omega)^d} + \frac{\lambda}{2}\|\nabla\times\mathbf{e}_{\mathbf{b}}\|^2_{L^2(\Omega)^d} - \Big(1 + \varrho^2\|\mathbf{u}\|^2_{L^\infty(\Omega)^d}\lambda^{-1}\Big)\|\mathbf{e}_{\mathbf{b}}\|^2_{L^2(\Omega)^d} \\ &+ |||(\mathbf{e}_{\mathbf{b}}, e_r)|||^2_h \ \le \ {S_1}^2 + 2{S_2}^2 + {S_3}^2. \end{aligned} \tag{14}$$

For $\mathbf{b} \in L^2(0,T;H^{k+1}(\Omega)^d)$ and $\partial_t \mathbf{b} \in L^2(0,T;H^k(\Omega))$, interpolation gives

$$S_1^2 = \|\partial_t \varepsilon_{\mathbf{b}}\|^2_{L^2(\Omega)^d} \le \; C \sum_K h_K^{2k} |\partial_t \mathbf{b}|^2_{H^k(\omega_K)}, \tag{15a}$$

$$S_2^2 \le C \sum_K h_K^{2k} \Big(\lambda + \varrho^2 \|\mathbf{u}\|^2_{L^\infty(K)} h_K^2 \lambda^{-1} \Big) |\mathbf{b}|^2_{H^{k+1}(\omega_K)}, \tag{15b}$$

$$S_3^2 \le C \sum_K h_K^{2k} \Big(\tau_{\mathbf{b}} d^2 + h_K^2 \tau_r^{-1} \Big) |\mathbf{b}|^2_{H^{k+1}(\omega_K)} \tag{15c}$$

where $\omega_K \in \Omega$ denotes an appropriate patch around cell K.

Now, (15c) suggests to set $\tau_r \tau_{\mathbf{b}} \sim h_K^2$ which with $\tau_r \sim \ell^2 \lambda^{-1}$ yields

$$\tau_r \sim \ell^2 \lambda^{-1}, \qquad \tau_{\mathbf{b}} \sim h_K^2 \lambda \ell^{-2}. \tag{16}$$

Moreover, from (15b) we observe boundedness of the parameter-dependent coefficient under a restriction on the mesh width h_K with

$$R_{mh} := \frac{\|\mathbf{u}\|_{L^\infty(K)} h_K}{\lambda} \le \frac{C}{\sqrt{\lambda}}. \tag{17}$$

Theorem 7 *Assume that* $\mathbf{b} \in L^2(0,T;[H^{k+1}(\Omega)]^d)$, $\partial_t \mathbf{b} \in L^2(0,T;[H^k(\Omega)]^d)$ *and* $r = 0$ *is a solution of* (3). *Moreover, let* $\mathbf{b}_h(0) = \mathbf{j}^{\mathbf{b}} \mathbf{b}(0)$. *Under the parameter choice* (16) *and mesh width restriction* (17)*, we obtain*

$$\begin{aligned} \varrho \|\mathbf{e_b}(t)\|^2_{L^2(\Omega)^d} + C \int_0^t e^{\tilde{C}_u(t-s)} \Big(\lambda \|\nabla \times \mathbf{e_b}(s)\|^2_{L^2(\Omega)^d} + |||(\mathbf{e_b}(s), e_r(s))|||_h^2 \Big) ds \\ \le \int_0^t e^{\tilde{C}_u(t-s)} C \sum_K h_K^{2k} \Big(|\partial_t \mathbf{b}(s)|^2_{H^k(\omega_K)} + |\mathbf{b}(s)|^2_{H^{k+1}(\omega_K)} \Big) ds. \end{aligned} \tag{18}$$

Proof Substituting (15) in (14), multiplication with $e^{-\tilde{C}_u t}$ with $\tilde{C}_u := 1 + \varrho^2 \|\mathbf{u}\|^2_{L^\infty(\Omega)^d} \lambda^{-1}$ and integration over (0,T) with an arbitrary $t \in (0,T]$ provides (18) since $\mathbf{e_b}(0) = \mathbf{j}^{\mathbf{b}} \mathbf{b}(0) - \mathbf{b}_h(0) = 0$. □

4 Numerical Simulations

While computations with the Taylor-Hood pair $\mathscr{P}_2/\mathscr{P}_1$ on triangular meshes were performed using FreeFem++ [6], we employed Gascoigne3d [4] for equal-order pair $\mathscr{Q}_1/\mathscr{Q}_1$ on quadrilateral meshes. Temporal discretizations performed by A-stable BDF2 and strongly A-stable fractional step theta scheme, respectively.

The parameter choice (16) refers to the worst case of singular solutions $\mathbf{b} \in V \cap H(\text{div } 0; \Omega)$ which do not belong to $H^1(\Omega)^d$. Numerical results for the flow around a re-entrant corner for equal-order pairs in [1] and for Taylor-Hood type pairs in [8] show that this choice is appropriate. The situation is slightly different for smooth solutions $\mathbf{b} \in V \cap H^1(\Omega)^d$. The PSPG-stabilization of the pseudo-pressure is still required for equal-order interpolation whereas it can be omitted for Taylor-Hood type elements.

Consider the flux expulsion phenomenon as magnetic field distortion in an infinitely long cylinder with radius $R = 0.5$ and cross-section $\Omega = (-0.5, 0.5)^2$. A magnetic field $\mathbf{b} = b_0\mathbf{e}_y$ pervades a conducting fluid in rigid body rotation (with given velocity $\mathbf{u} = (-y, x, 0), r < R$) inside the cylinder and the remainder being quiescent, i.e. $\mathbf{u} = (0, 0, 0), R < r$. Boundary $\partial\Omega$ is taken to be conducting. The exact solution for $\mathbf{b}$ can be found in [9] or [8].

The distortion of the field $\mathbf{b}$ becomes greater with increasing magnetic Reynolds number $R_m = \frac{\tilde{\omega}R^2}{\lambda}$ with $\tilde{\omega} = \frac{1}{2}(\nabla \times \mathbf{u}) = 1$. Then $\mathbf{b}$ is gradually expelled from the rotated fluid via combination of twisting of $\mathbf{b}$-lines and cross-stream diffusion which is related to the skin effect in conventional electromagnetism. In consequence of the dominance of convection over diffusion, we observe the formation of internal layers with width $\mathscr{O}(\sqrt{R_m})$.

We do not expect optimal convergence due to the discontinuity in the velocity field and the presence of internal layers. For magnetic Reynolds up to $R_m = 1000$, we found slightly better error rates for the $\mathscr{P}_2/\mathscr{P}_1$ pair compared to $\mathscr{Q}_1/\mathscr{Q}_1$, see Fig. 1. Moreover, we considered a further local projection stabilization of the Lorentz term $\nabla \times (\mathbf{u} \times (\varrho\mathbf{b}))$ for the equal-order case. The analysis of this paper can be easily extended to this case. For the $\mathscr{Q}_1/\mathscr{Q}_1$-pair this leads to improved results. Recent computations show that this stabilization is really required for even larger magnetic Reynolds numbers.

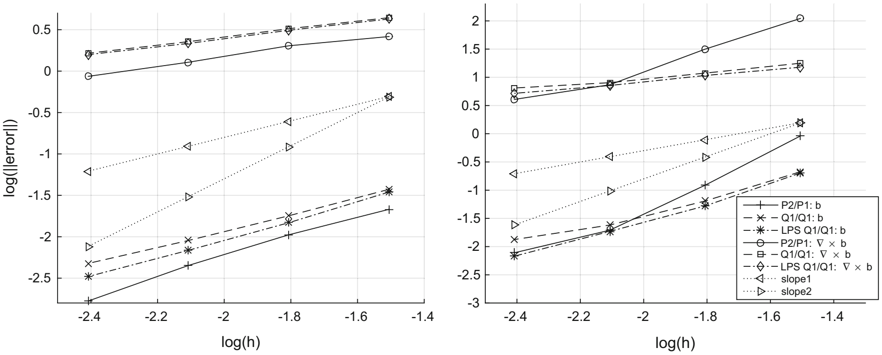


Fig. 1 Flux expulsion: convergence plots for $\|\mathbf{b} - \mathbf{b}_h\|_{L^2(\Omega)^d}$ and $\|\nabla \times \mathbf{b} - \mathbf{b}_h\|_{L^2(\Omega)^d}$ for $R_m = 10^2, T = 2500$ (*left*) and $R_m = 10^3, T = 10^6$ (*right*)

The problem of flux expulsion can be studied with more complicated examples in [9]. In [7], one can find further simulations of problems for given velocity fields **u** with different variants of eddies acting on the magnetic field.

5 Summary: Conclusions

We considered the application of Taylor-Hood type and equal order FE pairs for the magnetic field and the pseudo-pressure in the time-dependent magnetic induction problem. We gave a stability and convergence analysis for the spatially semi-discretized problem. In particular, we derived formulas for stabilization parameters of the divergence-free constraint of the magnetic field in a different way than in [1]. An additional local projection stabilization of the Lorentz term improved the results for increasing magnetic Reynolds numbers.

References

1. S. Badia, R. Codina, A nodal-based finite element approximation of the Maxwell problem suitable for singular solutions. SIAM J. Numer. Anal. **50**, 398–417 (2012)
2. S. Badia, R. Codina, R. Planas, Analysis of an unconditionally convergent stabilized finite element formulation for incompressible magnetohydrodynamics. Arch. Comput. Methods Eng. **22**, 621–636 (2015)
3. A. Bonito, J.-L. Guermond, Approximation of the eigenvalue problem for time harmonic Maxwell systems by continuous Lagrange finite elements. Math. Comput. **80**(276), 1887–1910 (2011)
4. M. Braack, R. Becker, T. Richter, B. Vexler, Gascoigne 3D – high performance adaptive finite element toolkit. http://www.gascoigne.de
5. M. Costabel, M. Dauge, Weighted regularization of Maxwell equations in polyhedral domains. Numer. Math. **93**(2), 239–278 (2002)
6. F. Hecht, New development in FreeFem++. J. Numer. Math. **20**(3–4), 251–265 (2012). 65Y15
7. U. Kaya, Numerical simulation of the induction equation using Lagrangian finite elements. Master Thesis, Georg-August University Göttingen, NAM (2014)
8. U. Kaya, B. Wacker, G. Lube, Stabilized nodal-based finite element methods for the magnetic induction problem. Math. Methods Appl. Sci. (2015). doi:0.1002/mma.3801
9. H.K. Moffatt, *Magnetic Field Generation in Electrically Conducting Fluids* (Cambridge University Press, Cambridge/New York, 1978)
10. P. Monk, *Finite Element Methods for Maxwell's Equations*. Numerical Mathematics and Scientific Computation (Clarendon Press, Oxford/New York, 2003)
11. G. Mur, Edge elements, their advantages and their disadvantages. IEEE Trans. Magn. **30**(5), 3552–3557 (1994)
12. E. Zeidler, L.F. Boron, *Nonlinear Functional Analysis and Its Applications. II/A: Linear Monotone Operators* (Springer, New York, 1989)

Modeling of a Three-Dimensional Spherulite Microstructure in Semicrystalline Polymers

H. Emre Oktay and Ercan Gürses

Abstract A finite element (FE) model, that explicitly discretizes a single 3D spherulite is proposed. A spherulite is a two-phase microstructure consisting of amorphous and crystalline regions. Crystalline regions, that grow from a central nucleus in the form of lamellae, have particular lattice orientations. In the FE analyses, 8-chain and crystal viscoplasticity constitutive models are employed. Stress-strain distributions and slip system activities in the spherulite microstructure are studied and found to be in good agreement with the literature. Influences of the crystallinity ratio on the yield stress and the initial Young's modulus are also investigated.

1 Introduction

Semicrystalline polymers (SCP) is a subset of polymers that stand out due to their toughness, high impact strength and wear resistance. Although some fundamental deformation mechanisms of SCPs are identified, a clear description of all stages of deformation is not available. Micromechanical multi-scale computational approaches may provide insight on the influence of deformation mechanisms taking place at the lower scales on the macroscopic response.

Segments of the polymer chains in SCPs orient in an orderly fashion to form crystalline lamellar structures. Unordered chains exist in SCPs as well, forming regions that are similar to the amorphous polymers. Crystalline lamellae could form complex structures such as spherulites where the lamellae grow from a central nucleus in radial directions. Between the lamellae, amorphous regions exist. Amorphous regions host tie chains that form interlamellar connections binding multiple lamellae together. As a result, amorphous and crystalline phases deform

H.E. Oktay
Department of Civil Engineering, Middle East Technical University, Ankara, Turkey
e-mail: emre.oktay@metu.edu.tr

E. Gürses (✉)
Department of Aerospace Engineering, Middle East Technical University, Ankara, Turkey
e-mail: gurses@metu.edu.tr

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_55

together and consistently. SCPs possess the deformation mechanisms of polymeric and crystalline materials. Crystalline phase of SCPs is orders of magnitude stiffer than the amorphous phase. Similar to the metals, dislocation based crystalline slip is the major plastic deformation mechanism of the crystalline phase. Amorphous phase on the other hand, is much easier to deform up to a locking stretch.

In literature homogenization techniques are used to obtain the macroscopic response of multi-phase materials. In these techniques, response of the constituent phases are generally represented by different constitutive equations. Additionally, to account for the crystalline texture, multiple crystalline grains are employed with orientations that are statistically representative of a crystalline texture. Then, the macroscopic response of the model is obtained by homogenization of the response of each individual grain via homogenization techniques [12, 16]. These studies do not explicitly take into account the spherulite morphology. Instead, randomly oriented aggregates are considered to be representative for the initially isotropic response of the spherulite morphology. Nevertheless, these models are advantageous for modeling SCPs having preoriented texture. One approach to consider the spherulitic morphology while employing the above models is solving representative volume elements (RVE) of spherulite with finite element method analysis [14, 15]. In these studies, due to the constitutive models employed at integration points, consistent deformation of amorphous and crystalline phases is accounted for at the microscale. However, at a higher scale level, consistent deformation of amorphous and crystalline phases within the spherulite is not addressed.

In this study we propose a three dimensional FEM discretization of a spherulite as an extension of our disk-like model [10]. The model allows consideration of arbitrary crystallinity ratios. Stress-strain distribution within the model is investigated for its compliance with the characteristic features of spherulite deformation. Heterogeneous slip system activity within the spherulite is compared with the literature. Influence of crystallinity ratio on the initial elastic modulus and the yield stress is reported.

2 Constitutive Models and Finite Element Model

Two constitutive models, employing finite deformation theory, are used in this study. A crystal viscoplasticity model is employed for the crystalline phase. The model utilizes the multiplicative decomposition of the deformation gradient $\mathbf{F} = \mathbf{F}^e\mathbf{F}^p$ into elastic and plastic contributions $\mathbf{F}^e$ and $\mathbf{F}^p$, respectively. Plastic flow is restricted to slip in given directions on predefined planes. Each one of these (direction, plane) pairs are referred to as a slip system. Rate of plastic deformation $\mathbf{L}^p$ is the sum of the slip rates on all slip systems, i.e., $\mathbf{L}^p = \dot{\mathbf{F}}^p\mathbf{F}^{p-1} = \sum_{\alpha=1}^{N} \dot{\gamma}^\alpha \mathbf{s}^\alpha \otimes \mathbf{m}^\alpha$. Here, $\mathbf{s}^\alpha$ and $\mathbf{m}^\alpha$ are defined in the reference configuration and are the slip direction and slip plane normal vectors, respectively. $\dot{\gamma}^\alpha$ is the slip rate of the slip system α. Evolution of the plastic component of deformation gradient is computed by utilizing the exponential integration scheme, i.e.,

$$\mathbf{F}^p_{n+1} = \Delta\mathbf{F}^p_{n+1}\mathbf{F}^p_n = \exp(\Delta t\textstyle\sum_{\alpha=1}^{N}\dot{\gamma}^\alpha_{n+1}\mathbf{s}^\alpha \otimes \mathbf{m}^\alpha)\mathbf{F}^p{}_n. \tag{1}$$

Tensor exponential is computed numerically according to the series expansion given in [5]. Stresses are calculated through Saint Venant-Kirchhoff Hyperelasticity $\mathbf{S} = \mathbb{C}^e : \mathbf{E}^e$. Here, $\mathbf{S}$ is the lattice based Second Piola-Kirchhoff stress, $\mathbb{C}^e$ is the fourth order elastic stiffness tensor and $\mathbf{E}^e = (\mathbf{F}^{e^T}\mathbf{F}^e - \mathbf{I})/2$ is the elastic Green-Lagrange strain tensor. Evolution of the slip rates $\dot{\gamma}^\alpha$ are defined by a power law type constitutive equation

$$\dot{\gamma}^\alpha = \dot{\gamma}_0 \left(\tau^\alpha/\tau^\alpha_y\right) \left|\tau^\alpha/\tau^\alpha_y\right|^{n-1} \tag{2}$$

where $\dot{\gamma}_0$, τ^α_y, n are the reference shear rate, the critical resolved shear stress and the rate sensitivity parameter, respectively. $\tau^\alpha = \mathbf{R}^{e^T}\boldsymbol{\tau}\mathbf{R}^e : (\mathbf{s}^\alpha \otimes \mathbf{m}^\alpha)$ is the Schmid shear stress [5]. $\mathbf{R}^e$ is the rotation tensor and $\boldsymbol{\tau} = \mathbf{F}^e\mathbf{S}\mathbf{F}^{e-T}$ is the lattice based Kirchhoff stress. Material parameters of the model are presented in Table 1. Slip systems and τ^α_y values are given in Table 2. The 8-chain [2] rubber elasticity model is used for the amorphous phase. Employed values of the amorphous phase material parameters; bulk modulus $\kappa = 2$ GPa, shear modulus $\mu = 35$ MPa and locking stretch $\lambda_{lock} = 7$ are taken from [16] for the high density polyethylene (HDPE).

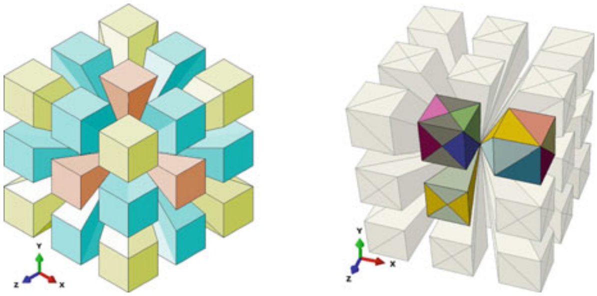
A spherulite FEM model is constructed by dividing a cube into amorphous and crystalline regions. Crystalline phase regions (lamellae) originate from the nucleus as shown in Fig. 1a. Lamellae could be grouped into three according to their shapes. There are total 6, 8 and 12 lamellae in the lamella group FC, C and ME respectively. Lamellae of FC, C and ME are directed towards the face centers, corners and mid points of the edges of the cube, respectively. Regardless of the lamellae group they belong, edge lengths of the pyramid bases are of equal length. In this study Polyethylene (PE) spherulite is considered. Therefore, orientation of crystallographic axes comply with the following criteria: (i) crystal lamella growth

Table 1 Material parameters employed in crystalline phase material model

Elastic constants for PE crystal [3]						Viscoplasticity parameters [16]		
$\mathbb{C}_{11}$	$\mathbb{C}_{33}$	$\mathbb{C}_{12}$	$\mathbb{C}_{13}$	$\mathbb{C}_{44}$	$\mathbb{C}_{66}$	$\dot{\gamma}_0$	n	τ_0
[GPa]	[GPa]	[GPa]	[GPa]	[GPa]	[GPa]	[s^{-1}]		[MPa]
7	81	3.8	4.7	1.5	1.6	1×10^{-3}	9	8

Table 2 Slip systems of polyethylene crystal [8]

	Slip system	τ^α_y/τ_0		Slip system	τ^α_y/τ_0
Chain slip	(100)[001]	1.0	Transverse slip	(100)[010]	1.66
	(010)[001]	2.5		(010)[100]	2.5
	{110}[001]	2.5		$\{110\}\langle 1\bar{1}0\rangle$	2.2



(a) Red: lamellae FC, Yellow: lamellae C, Blue: lamellae ME

(b) Regions with the same coordinate systems are shown in same colors.

Fig. 1 (**a**) Categorization of lamellae of the spherulite model according to lamellae geometry, (**b**) typical regions of each lamellae group where the same coordinate system is employed for crystallographic orientation

is in the b lattice axis direction [9], (ii) interface between the amorphous phase and the lamellae is $\{201\}$ crystallographic plane [6]. For the first condition a spherical coordinate system, origin of which is located at the nucleus is employed. Then, b lattice axis is oriented in the direction of the radial basis vector e_r. Note that any lamella of our model has multiple large planar surfaces as shown in Fig. 1a that form interfaces with the amorphous phase. In this study it is decided to satisfy the condition (ii) for each interface of a lamella. To this end, each lamellae is divided into regions according to the interface plane they belong and different orientations are assigned to each region to satisfy the condition (ii). Lamellae belonging to groups FC, C and ME are divided into 2, 6 and 5 regions respectively shown in different colors in Fig. 1b. e_θ basis vector of the spherical coordinate system is selected to define the orientation of the chain. For each lamella region, orientation of the coordinate system is selected such that interface plane normal vector lies in the plane spanned by e_r and e_θ. Finally, to set the desired angle between the interface normal and the chain direction, the coordinate system is rotated 35° around the b axis.

3 Results

Strain distribution within the spherulite strongly depends on the orientation of lamellae with respect to the loading direction. In literature, spherulite is divided into three regions according to the lamella orientation with respect to the tensile loading direction [9]. These regions, referred to as the equatorial, polar and inclined regions, are defined as follows: In polar regions, lamella growth direction is nearly parallel

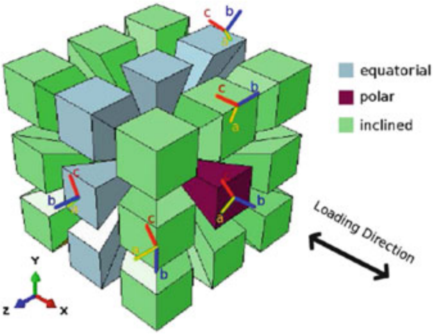


Fig. 2 Equatorial, polar and inclined regions of spherulite and the representative lattice directions for the tensile loading along x direction

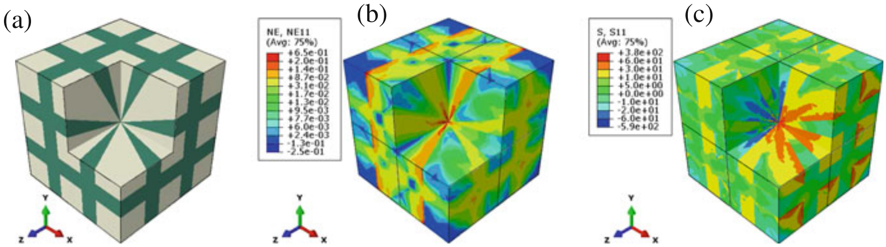


Fig. 3 Spatial variation of stress and strain components at 3 % macroscopic engineering strain. 1/8th of the spherulite is hidden for visualization. In (**a**) dark and light colored regions are amorphous and crystalline, respectively. (**a**) Two different phases. (**b**) Nominal strain in loading direction. (**c**) Cauchy stress along the radial direction

to the loading direction. In equatorial regions lamellae growth direction is nearly perpendicular to the loading direction. Regions that neither belong to the equatorial nor to the polar regions belong to the inclined region, see Fig. 2. Figure 3 presents the spatial variation of stress and strain components within the spherulite at 3 % engineering strain. In Fig. 3 it is seen that stress and strain depend on the distance from the nucleus, while the maximum values are at the nucleus. Radial stresses in equatorial regions are compressive. After the nucleus, maximum strain occurs at the amorphous phase of equatorial regions showing interlamellar separation, indicating that deformation begins at nucleus and spreads along the equatorial regions along the radius. All of these observations are in agreement with the important features of spherulite deformation presented in [11].

Slip system activities at the engineering strain of 0.163 are reported below. As the measure of activity, accumulated plastic slip $\sum_{t=0} |\Delta t \dot{\gamma}^{\alpha}|$ is employed for each slip system, where Δt is the time increment. In Table 3 relative activities of each slip system throughout the microstructure are presented. Relative activity of a slip system is the ratio of the sum of accumulated slip of that system to the sum of accumulated slip of all slip systems at all integration points of the model.

Table 3 Relative activities of slip systems within the equatorial, polar, inclined regions and the complete spherulite between 0 and 0.163 engineering strain levels

	(100)[001]	(010)[001]	(110)[001]	$(1\bar{1}0)[001]$	(100)[010]	(010)[100]	$(110)[1\bar{1}0]$	$(1\bar{1}0)[110]$
Equatorial (%)	73.9	0.6	0.6	1.0	0.9	0.0	11.3	11.7
Polar (%)	9.3	0.6	0.5	0.5	2.2	0.1	43.8	43.0
Inclined (%)	19.0	10.1	5.5	5.6	41.8	1.0	8.4	8.6
Complete spherulite (%)	34.0	3.8	2.2	2.4	15.0	0.4	21.1	21.1

Table 3 also presents the relative activities of each slip system within equatorial, polar and inclined regions separately. Considering the complete spherulite, from Table 3 it is seen that $\{110\}\langle 1\bar{1}0\rangle$, (100)[001] and (100)[010] are the most active slip systems. These systems are the dominant slip systems of polar, equatorial and inclined regions, respectively. In inclined regions, all slip systems are active. In equatorial regions only the (100)[001] and $\{110\}\langle 1\bar{1}0\rangle$ systems have considerable activity. Finally in polar regions (100)[001], (100)[010] and $\{110\}\langle 1\bar{1}0\rangle$ systems have activity.

From Fig. 4a it is seen that (100)[001] slip system is active in the equatorial and some of the inclined regions. In the polar region, there is a slight activity. In Fig. 4b, (010)[001] slip system is seen to have almost no activity in the polar and equatorial regions, while it is active in the inclined regions. In Fig. 4e it is seen that (100)[010] slip system is active and dominant only in the inclined regions. From Fig. 4f it is seen that (010)[100] has negligible activity throughout the spherulite. Since (010)[100] and (100)[010] slip systems are orthogonal to each other, observed activity difference between these systems is only due to the difference between their slip resistances. According to Fig. 4g, h, $\{110\}\langle 1\bar{1}0\rangle$ slip systems have activity in all lamellae. High activity of (100)[001] chain slip in equatorial regions are in agreement with the expectations of [7], as it facilitates the chain alignment to the loading direction. Activity of $\{110\}\langle 1\bar{1}0\rangle$ slip system in equatorial regions is in agreement with [1] where slip on $\{110\}\langle 1\bar{1}0\rangle$ is expected in equatorial regions. In summary, two slip systems are observed to be active in equatorial regions in our study: (100)[001] and $\{110\}\langle 1\bar{1}0\rangle$. Activity of (100)[001] and $\{110\}\langle 1\bar{1}0\rangle$ is in harmony with [7] and [1], respectively. In inclined regions [1] expects chain slip and slip on the $\{110\}$ planes. In inclined regions, all chain slip systems are active in our study. In the polar regions [1] expects $\{110\}\langle 1\bar{1}0\rangle$ as the major slip system parallel to our finding. Finally, the effect of the crystallinity is studied. Figure 5 presents the comparison of the yield stress and the Young's modulus obtained from this study with the experimental study of [4]. It is seen that the predictions of this study, especially the slopes of change of the yield stress and the Young's modulus are in good agreement with [4].

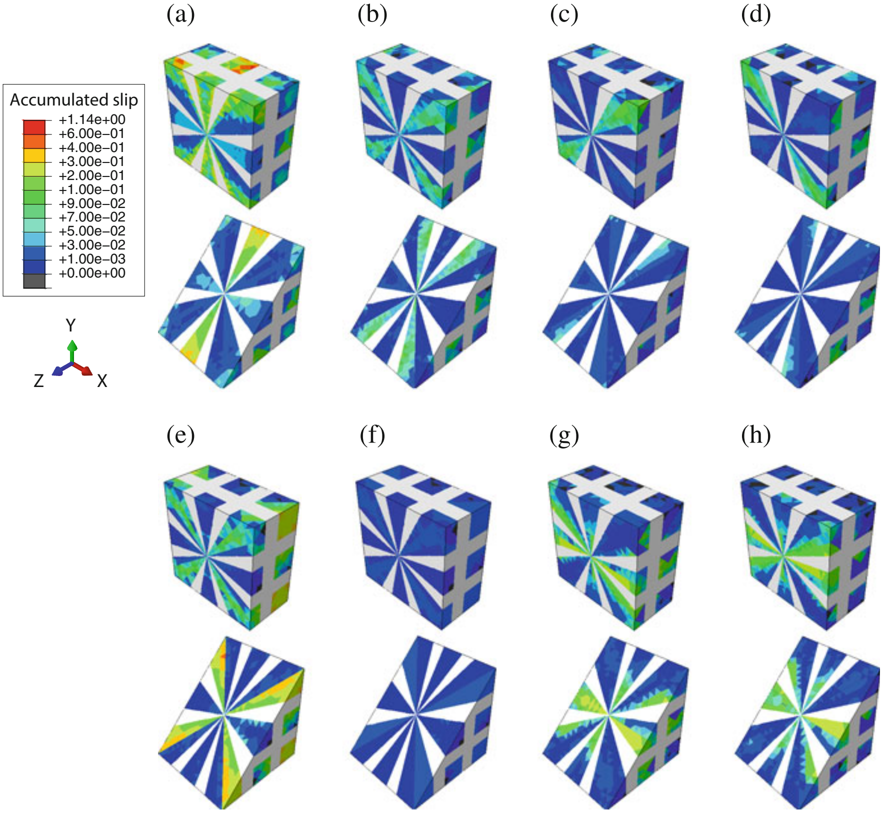


Fig. 4 Distribution of slip amounts shown for a vertical cut and an inclined cut with a normal vector of (0,1,1). Loading is in the x-direction. (**a**) (100)[001]. (**b**) (010)[001]. (**c**) (110)[001]. (**d**) $(1\bar{1}0)$[001]. (**e**) (100)[010]. (**f**) (010)[100]. (**g**) $(110)[1\bar{1}0]$. (**h**) $(1\bar{1}0)[110]$

3D finite element model of a single spherulite is studied. Stress and strain distribution characteristics of the model indicate that the model possesses the important features of the spherulite deformation. Slip system activities throughout the spherulite are reported. Influences of crystallinity ratio on the yield stress and the initial elastic modulus are investigated. A general good agreement with the literature is found. It should be noted that this study considers a single spherulite. On the other hand, the study of [4] is on macroscopic samples that are composed of multiple spherulites. As discussed in [7] and also recently presented in [13]; the macroscopic strain, in general, is accommodated by not a single spherulite but a collection of spherulites. Therefore, a multi-spherulitic model may be more representative for a comparison with the results of [4].

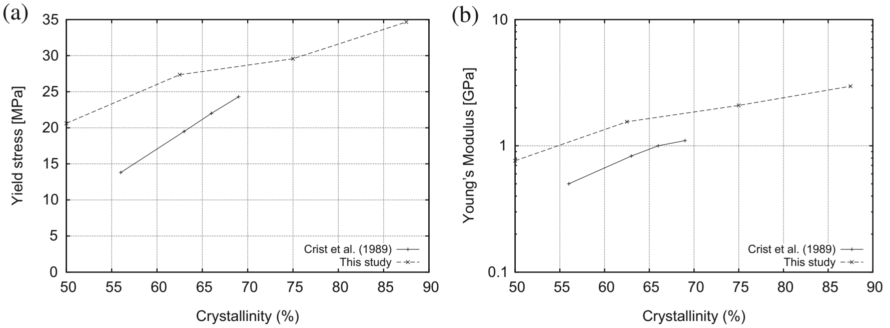


Fig. 5 Comparison of findings of this study with experimental study of [4] on the variation of macroscopic yield stress and Young's modulus with crystallinity ratio. (**a**) Yield stress. (**b**) Young's modulus

Acknowledgements This work was supported by the Scientific and Technological Research Council of Turkey (TÜBİTAK), Grant No. 111M646.

References

1. P. Allan, M. Bevis, Deformation processes in thin melt-cast films of high-density polyethylene – 2. deformation processes in the non-equatorial regions of spherulites. Philos. Mag. A **41**(4), 555–572 (1980)
2. E.M. Arruda, M.C. Boyce, A three-dimensional constitutive model for the large stretch behavior of rubber elastic materials. J. Mech. Phys. Solids **41**(2), 389–412 (1993)
3. C.L. Choy, W.P. Leung, Elastic moduli of ultradrawn polyethylene. J. Polym. Sci. A-2, Polym. Phys. **23**(9), 1759–1780 (1985)
4. B. Crist, C.J. Fisher, P.R. Howard, Mechanical properties of model polyethylenes: tensile elastic modulus and yield stress. Macromolecules **22**(5), 1709–1718 (1989)
5. E.A. de Souza Neto, D. Peric, D.R.J. Owen, *Computational Methods for Plasticity: Theory and Applications* (Wiley, West Sussex, 2008)
6. S. Gautam, S. Balijepalli, G.C. Rutledge, Molecular simulations of the interlamellar phase in polymers: effect of chain tilt. Macromolecules **33**(24), 9136–9145 (2000)
7. I.L. Hay, A. Keller, Polymer deformation in terms of spherulites. Kolloid-Z. Z. für Polym. **204**(1–2), 43–74 (1965)
8. B.J. Lee, A.S. Argon, D.M. Parks, S. Ahzi, Z. Bartczak, Simulation of large strain plastic deformation and texture evolution in high density polyethylene. Polymer **34**(17), 3555–3575 (1993)
9. L. Lin, A.S. Argon, Structure and plastic deformation of polyethylene. J. Mater. Sci. **29**(2), 294–323 (1994)
10. H.E. Oktay, E. Gürses, Modeling of spherulite microstructures in semicrystalline polymers. Mech. Mater **90**, 83–101 (2015)
11. E.F. Oleinik, Plasticity of semicrystalline flexible-chain polymers at the microscopic and mesoscopic levels. Polym. Sci. Ser. C **45**(1), 17–117 (2003)
12. D.M. Parks, S. Ahzi, Polycrystalline plastic deformation and texture evolution for crystals lacking five independent slip systems. J. Mech. Phys. Solids **38**(5) 701–724 (1990)

13. J. Teixeira-Pinto, C. Nadot-Martin, F. Touchard, M. Gueguen, S. Castagnet, Towards the size estimation of a representative elementary domain in semi-crystalline polymers. Mech. Mater. **95**, 239–247 (2016)
14. M. Uchida, N. Tada, Micro-, meso- to macroscopic modeling of deformation behavior of semicrystalline polymer. Int. J. Plast. **49**, 164–184 (2013)
15. J.A.W. Van Dommelen, D.M. Parks, M.C. Boyce, W.A.M. Brekelmans, F.P.T. Baaijens, Micromechanical modeling of intraspherulitic deformation of semicrystalline polymers. Polymer **44**(19), 6089–6101 (2003)
16. J.A.W. Van Dommelen, D.M. Parks, M.C. Boyce, W.A.M. Brekelmans, F.P.T. Baaijens, Micromechanical modeling of the elasto-viscoplastic behavior of semi-crystalline polymers. J. Mech. Phys. Solids **51**(3), 519–541 (2003)

Numerical Approximation of Interaction of Fluid Flow and Elastic Structure Vibrations

Jan Valášek, Petr Sváček, and Jaromír Horáček

Abstract This paper deals with flow induced vibrations of an elastic body. A simplified model of the human vocal fold is mathematically described. In order to consider the time dependent domain the arbitrary Lagrangian-Eulerian method is used. The viscous incompressible fluid flow and linear elasticity models are considered. The developed numerical schemes for the fluid flow and the elastic body are implemented by the in-house developed solver based on the finite element method. Preliminary numerical results testing the convergence of solver are presented.

1 Introduction

The problem of interaction of fluid flow and elastic structure is widely spread in nature and it has important applications not only in technical practice. Beside well-known bridge oscillations in wind or stability of an airfoil in fluid flow, see [3], the biological applications are newly investigated, as e.g. blood flow in arteries or flow induced vibrations of vocal folds, see e.g. [8, 10] or [12].

The coupled problem of fluid-structure interaction (FSI) can be solved by many different approaches but in most of them the arbitrary Lagrangian Eulerian (ALE) method for description of fluid flow in time dependent domain is used, see [12]. In this paper the considered mathematical model is presented and the discretization by the finite element method (FEM) is described. The fluid and structural problems are solved by specific solvers on each domain and coupled via boundary conditions on the common interface. This partitioned scheme is strongly coupled.

J. Valášek (✉) • P. Sváček
Faculty of Mechanical Engineering, CTU, Karlovo nám. 13, 121 35 Praha 2, Czech Republic
e-mail: valasek.jan@volny.cz; petr.svacek@fs.cvut.cz

J. Horáček
Institute of Thermomechanics, Academy of Sciences of the Czech Republic, Dolejškova 5, 182 00 Praha 8, Czech Republic
e-mail: jaromirh@it.cas.cz

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_56

The numerical results are presented for a simple test case, where the convergence of the numerical solution in time and space steps are tested.

2 Mathematical Model

For the sake of simplicity a two dimensional problem – shown in Fig. 1 – is considered. Here, domain Ω^s_{ref} denotes the reference representation of an elastic structure, Ω^f_{ref} is the reference domain occupied by the fluid at the time instant $t = 0$ and $\Gamma_{\mathrm{W_{ref}}} = \Gamma_{\mathrm{W_0}}$ is the common interface. At time instant t the fluid domain Ω^f_{ref} turns to Ω^f_t. The deformation of domain $\Omega^s = \Omega^s_{ref}$ at time t is handled by the Lagrange approach, i.e. in the reference coordinates.

2.1 Elastic Structure

The motion of elastic body Ω^s_{ref} is described by the partial differential equation expressing dynamical equilibrium between inertia force and the applied surface and volume forces

$$\rho^s \frac{\partial^2 \mathbf{u}}{\partial t^2} - \frac{\partial \tau^s_{ij}}{\partial x_j} = \mathbf{f}^s \quad \text{in } \Omega^s \times (0, \mathrm{T}), \tag{1}$$

where the vector $\mathbf{u}(x, t)$ denotes the displacement vector, $\mathbf{f}^s$ is the volume force, ρ^s is the structure density and τ_{ij} are the components of the Cauchy stress tensor. These components are for the isotropic Hooke's material and small displacements given by

$$\tau^s_{ij} = \lambda^s \mathrm{div}\, \mathbf{u}\, \delta_{ij} + 2\mu^s e^s_{ij}(\mathbf{u}), \tag{2}$$

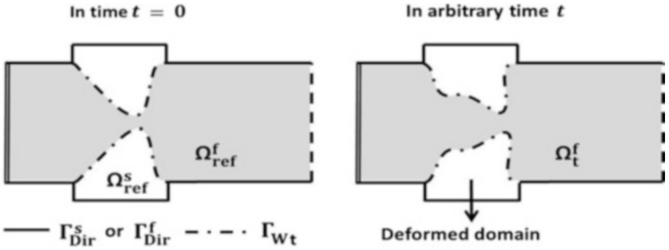


Fig. 1 Scheme of vocal folds model with boundaries marked before and after deformation

where δ_{ij} is Kronecker's delta and $e^s_{jk}(\mathbf{u}) = \frac{1}{2}\left(\frac{\partial u_j}{\partial x_k} + \frac{\partial u_k}{\partial x_j}\right)$ is the strain tensor with the Lame's constants λ^s, μ^s. Problem (1) is completed with the given initial and boundary conditions

$$\begin{aligned}
&\text{(a)} && \mathbf{u}(X,0) = \mathbf{u}_0(X), && \text{for } X \in \ \Omega^s,\\
&\text{(b)} && \frac{\partial \mathbf{u}}{\partial t}(X,0) = \mathbf{u}_1(X) && \text{for } X \in \ \Omega^s, \qquad (3)\\
&\text{(c)} && \mathbf{u}(X,t) = \mathbf{u}_{\mathrm{Dir}}(X,t) && \text{for } X \in \Gamma^s_{\mathrm{Dir}},\ t \in (0,\mathrm{T}),\\
&\text{(d)} && \tau^s_{ij}(X,t)\, n^s_j(X) = q^s_i(X,t), && \text{for } X \in \Gamma^s_{\mathrm{W_t}},\ t \in (0,\mathrm{T}),
\end{aligned}$$

where the $\Gamma_{\mathrm{W_t}}, \Gamma^s_{\mathrm{Dir}}$ are mutually disjoint parts of the boundary $\partial\Omega = \Gamma_{\mathrm{W_t}} \cup \Gamma^s_{\mathrm{Dir}}$ (see Fig. 1) and $n^s_j(X)$ are components of outer unit normal to $\Gamma_{\mathrm{W_t}}$.

2.2 ALE Method

In order to address the time discretization on time dependent domain Ω^f_t the ALE method is used. This method is based on a difeomorphic mapping A_t of any reference point $X \in \Omega^f_{ref}$ on the point of deformed domain $x = A_t(X) \in \Omega^f_t$. The ALE domain velocity $\mathbf{w}_D$, i.e. the velocity of a point with the given reference X, is defined by

$$\mathbf{w}_D(x,t) = \hat{\mathbf{w}}_D(A_t^{-1}(x),t), \quad t \in (0,\mathrm{T}),\ x \in \Omega^f_t, \tag{4}$$

where $\hat{\mathbf{w}}_D(X,t) = \frac{\partial}{\partial t}A_t(X)$, for $t \in (0,\mathrm{T})$ and $X \in \Omega^f_{ref}$. The ALE derivative, i.e. the time derivative with respect to a fixed reference X, can be expressed as (see e.g. [9])

$$\frac{D^A}{Dt}f(x,t) = \frac{\partial f}{\partial t}(x,t) + \mathbf{w}_D(x,t)\cdot\nabla f(x,t). \tag{5}$$

2.3 Fluid Flow

The motion of the viscous incompressible fluid in a time dependent domain Ω^f_t is modelled by the Navier-Stokes equations written in the ALE form

$$\frac{D^A\mathbf{v}}{Dt} + ((\mathbf{v} - \mathbf{w}_D)\cdot\nabla)\mathbf{v} - \nu^f\Delta\mathbf{v} + \nabla p = \mathbf{0}, \quad \operatorname{div}\mathbf{v} = 0 \quad \text{in } \Omega^f_t, \tag{6}$$

where $\mathbf{v}(x,t)$ denotes the fluid velocity, p is the kinematic pressure and ν^f is the kinematic fluid viscosity.

The system of Eqs. (6) is equipped with zero initial and the following boundary conditions

$$\begin{aligned} &\text{(a)} & \mathbf{v}(x,t) &= \mathbf{w}_D(x,t) && \text{for } x \in \Gamma^f_{\mathrm{Dir}} \cup \Gamma_{\mathrm{W_t}},\ t \in (0,\mathrm{T}), \\ &\text{(b)} & \mathbf{v}(x,t) &= \mathbf{v}_{\mathrm{Dir}}(x,t) && \text{for } x \in \Gamma^f_{\mathrm{In}},\ t \in (0,\mathrm{T}), \\ &\text{(c)} & p(x,t)\mathbf{n}^f - \nu^f \frac{\partial \mathbf{v}}{\partial \mathbf{n}^f}(x,t) &= -\frac{1}{2}\mathbf{v}(\mathbf{v}\cdot\mathbf{n}^f)^- && \text{for } x \in \Gamma^f_{\mathrm{Out}},\ t \in (0,\mathrm{T}), \end{aligned} \tag{7}$$

where $\mathbf{n}^f$ here denotes outer unit normal to boundary Γ^f_{Out}. The last condition (7 c) is the modified do-nothing boundary condition according to [2], which increases the stability of the scheme and suppresses backward inlet through the outlet boundary.

2.4 Coupling Conditions

The FSI problem is solved by the partitioned approach. This means that the elastic structure and the fluid flow problem is approximated by different solvers and coupled together with the aid of the interface boundary conditions prescribed on the interface $\Gamma_{\mathrm{W_t}}$. Let us mention here that the location of the interface is also unknown at each time instant t and depends on the establishing force equilibrium between the aerodynamic and elastic forces. Thus $\Gamma_{\mathrm{W_t}}$ depends on the displacement $\mathbf{u}$ at time t by

$$\Gamma_{\mathrm{W_t}} = \left\{x \in \mathbb{R}^2 | \, x = X + \mathbf{u}(X,t),\ X \in \Gamma_{\mathrm{W_{ref}}}\right\}. \tag{8}$$

The force equilibrium at $\Gamma_{\mathrm{W_t}}$ leads to the dynamic condition prescribed in problem (1) by boundary condition (3 d), where

$$q_i^s(X,t) = \sum_{j=1}^{2} \rho^f \left(p\delta_{ij} - \nu^f \left(\frac{\partial \mathbf{v}_i}{\partial x_j} + \frac{\partial \mathbf{v}_j}{\partial x_i}\right)\right) n_j^f(x), \tag{9}$$

and where n_j^f denotes the components of the unit normal (here to interface $\Gamma_{\mathrm{W_t}}$) oriented out of Ω_t^f at $x = X + \mathbf{u}(X,t),\ X \in \Gamma_{\mathrm{W_{ref}}}$.

Furthermore on the interface $\Gamma_{\mathrm{W_t}}$ the kinematic condition (7a) is prescribed.

3 Numerical Model

The FSI problem is now discretized in space using the FEM. For the time discretization of (1) and (6) the Newmark scheme and the backward differentiation formula of second order (BDF2) are used, respectively. The time interval [0, T] is divided equidistantly, i.e. $t_n = n\Delta t$.

3.1 Elastic Structure

First, Eq. (1) is reformulated in the weak sense by multiplying of Eq. (1) by a test function $\boldsymbol{\varphi} \in \mathbf{V}$, integration over Ω^s, using the Green's theorem and Hooke's law (2). It means that we seek $\mathbf{u} \in \mathbf{V}$ such that

$$\left(\rho^s \frac{\partial^2 \mathbf{u}}{\partial t^2}, \boldsymbol{\varphi}\right)_{\Omega^s} + \left(\lambda^s (\operatorname{div} \mathbf{u})\, \delta_{ij} + 2\mu^s e^s_{ij}(\mathbf{u}), e^s_{ij}(\boldsymbol{\varphi})\right)_{\Omega^s} = (\mathbf{f}^s, \boldsymbol{\varphi})_{\Omega^s} + (\mathbf{q}^s, \boldsymbol{\varphi})_{\Gamma^s_{\text{Neu}}}, \tag{10}$$

holds for all $\boldsymbol{\varphi} \in \mathbf{V}$. Here $(\cdot, \cdot)_{\Omega^s}$ denotes scalar product in the space $L^2(\Omega^s)$ and $\boldsymbol{L}^2(\Omega^s)$, the space $\mathbf{V} = V \times V$, where $V = \{\phi \in W^{1,2}(\Omega^s) | \phi = 0 \text{ on } \Gamma^s_{\text{Dir}}\}$, and $W^{1,2}(\Omega)$ is the Sobolev's space, see [1].

Replacing the space $\mathbf{V}$ by it's subspace $\mathbf{V}_h \subset \mathbf{V}$, the solution is sought in the form $\mathbf{u}_h(x, t) = \sum_{j=1}^{N_h} \alpha_j(t) \boldsymbol{\varphi}_j(x)$, where functions $\boldsymbol{\varphi}_j$ form a base of $\mathbf{V}_h$ and $N_h = \dim V_h$. Then Eq. (10) can be written in the matrix form

$$\mathbb{M}^T \ddot{\boldsymbol{\alpha}} + \mathbb{K}^T \boldsymbol{\alpha} = \mathbf{b}(t), \tag{11}$$

where the vector $\mathbf{b}(t) = (b_i(t)) \implies b_i(t) = (\mathbf{f}^s, \boldsymbol{\varphi}_i)_{\Omega^s} + (\mathbf{q}^s, \boldsymbol{\varphi}_i)_{\Gamma^s_{\text{Neu}}}$ and the elements of the matrices $\mathbb{M} = (m_{ij}), \mathbb{K} = (k_{ij})$ are given by

$$m_{ij} = (\rho^s \boldsymbol{\varphi}_j, \boldsymbol{\varphi}_i)_{\Omega^s}, \quad k_{ij} = (\lambda^s (\operatorname{div} \boldsymbol{\varphi}_j)\, \delta_{rl} + 2\mu^s e^s_{rl}(\boldsymbol{\varphi}_j), e^s_{rl}(\boldsymbol{\varphi}_i))_{\Omega^s},$$

The resulting system of second order ordinary differential Eqs. (11) is solved by Newmark method, see [4].

3.2 Fluid Flow

In order to discretize Eq. (6) in time the BDF2 scheme is applied, so the ALE derivative is approximated by

$$\frac{D^A \mathbf{v}}{Dt}(t_{n+1}) \approx \frac{3\mathbf{v}^{n+1} - 4\overline{\mathbf{v}}^n + \overline{\mathbf{v}}^{n-1}}{2\Delta t}, \tag{12}$$

where for a fixed time instant t_{n+1} we denote $\overline{\mathbf{v}}^i(x) = \mathbf{v}^i(\tilde{x})$ for $\tilde{x} = A_{t_i}(A^{-1}_{t_{n+1}}(x))$, $i \in \{n-1, n\}$ and $x \in \Omega^f_{n+1}$. For the sake of simplicity in what follows the time index ${}^{n+1}$ shall be omitted, e.g. $\Omega^f := \Omega^f_{t_{n+1}}$.

The application of BDF2 in Eq. (6) and the standard derivation of the weak formulation yields

$$\left(\frac{3\mathbf{v} - 4\overline{\mathbf{v}}^n + \overline{\mathbf{v}}^{n-1}}{2\Delta t}, \boldsymbol{\Phi}\right)_{\Omega^f} + c(\mathbf{v};\mathbf{v};\boldsymbol{\Phi}) + \frac{1}{2}((\mathbf{v}\cdot\mathbf{n})^+\mathbf{v}, \boldsymbol{\Phi})_{\Gamma^f_{\mathrm{Out}}} + \\ + \nu^f(\nabla\mathbf{v}, \nabla\boldsymbol{\Phi})_{\Omega^f} - (p, \nabla\boldsymbol{\Phi})_{\Omega^f} + (q, \operatorname{div}\mathbf{v})_{\Omega^f} = 0, \qquad (13)$$

which should be satisfied for any test function $\boldsymbol{\Phi}$ from the space $\mathbf{X} = X \times X$, $X = \{f \in W^{1,2}(\Omega^f) | f = 0 \text{ on } \Gamma^f_{\mathrm{Dir}} \cup \Gamma^f_{\mathrm{In}} \cup \Gamma^f_{\mathrm{W}_{\mathrm{t}_{n+1}}}\} \subset W^{1,2}(\Omega^f)$ and any $q \in M = L^2(\Omega^f)$. Here, the term $c(\cdot,\cdot,\cdot)$ is the form defined by

$$c^{n+1}(\mathbf{z};\mathbf{v};\boldsymbol{\Phi}) = \frac{1}{2}(((\mathbf{z} - 2\mathbf{w}^{n+1}_D)\cdot\nabla)\mathbf{v}, \boldsymbol{\Phi})_{\Omega^f} - \frac{1}{2}((\mathbf{z}\cdot\nabla)\boldsymbol{\Phi}, \mathbf{v})_{\Omega^f}. \qquad (14)$$

During derivation of this scheme the nonlinear boundary condition (7c) naturally arises, see [2].

Then by FEM we approximate spaces $\mathbf{X}$ and M by the finite dimensional spaces $\mathbf{X}_h$ and M_h, so the solution $\mathbf{v} \approx \mathbf{v}_h$ can be expressed as linear combination of basis functions leading to the system

$$\begin{pmatrix} \mathbb{A}(\mathbf{v}^*_h) & \mathbb{B} \\ \mathbb{B}^T & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{pmatrix} = \begin{pmatrix} \mathbf{g} \\ \mathbf{0} \end{pmatrix}, \qquad (15)$$

where $\boldsymbol{\beta}, \boldsymbol{\gamma}$ are vectors of linear combination coefficients, $\mathbb{A}(\mathbf{v}^*_h) = \frac{1}{\Delta t}\mathbb{M} + \mathbb{C}(\mathbf{v}^*_h) + \mathbb{D}$ and the elements of the matrices $\mathbb{M} = (m_{ij}), \mathbb{C} = (c_{ij}), \mathbb{D} = (d_{ij})$ and vector $\mathbf{g} = (g_i)$ are given by

$$m_{ij} = \frac{3}{2}(\boldsymbol{\Phi}_j, \boldsymbol{\Phi}_i)_{\Omega^f}, \;\; c_{ij} = c^{n+1}(\mathbf{v}^*_h; \boldsymbol{\Phi}_j; \boldsymbol{\Phi}_i), \;\; d_{ij} = \nu^f(\nabla\boldsymbol{\Phi}_j, \nabla\boldsymbol{\Phi}_i)_{\Omega^f}, \\ b_{ij} = (-q_j, \operatorname{div}\boldsymbol{\Phi}_i)_{\Omega^f}, \quad g_i = (\frac{4\overline{\mathbf{u}}^n - \overline{\mathbf{u}}^{n-1}}{2\Delta t}, \boldsymbol{\Phi}_i)_{\Omega^f}. \qquad (16)$$

The nonlinear system of Eqs. (15) is solved by the linearization $\mathbf{v}^*_h = \mathbf{v}^n$ and then the mathematical library UMFPACK is employed, see [5]. The P1-bubble/P1 elements are used in the numerical simulations, which according to [6] satisfy the well-known Babuška-Brezzi condition.

4 Numerical Results

All numerical simulation were performed on the mesh of vocal fold model M5 suggested by paper [7] and shown in Fig. 2. Here, only one half of the channel with assumption of solution symmetry was used due to reduction of computation cost and verification of the preliminary results.

All numerical results were achieved for flow driven vibrations from reference state Ω^s_{ref} induced by parabolic inlet profile with maximum $v_1 = 0.15\,\mathrm{m/s}$, $\nu^f = 1.5 \cdot 10^{-5}\,\mathrm{m^2/s}$ and $\rho^f = 1.7\,\mathrm{kg/m^3}$. The material properties are the same as in [11]. Examples of flow field at two time instants is shown in Fig. 3. The structure deformation was enabled after $t^S = 0.01$ s, where the fluid field is fully developed. Although the start of simulation is unphysical – sudden release of interaction, after a short time (0.15 s) periodic oscillations around a deflected position arose.

The solution on three meshes (a coarse, once refined and twice refined mesh) for $\Delta t = 2 \cdot 10^{-4} s$ was performed. Figure 4 shows the displacement of the point A in direction x and it's Fourier transformation for different meshes. The time signals are very similar, small difference can be seen in the frequency spectra, where the first dominant frequency is shifted by 7 Hz because of shorter time signal.

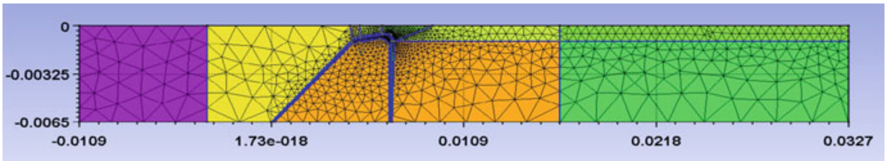


Fig. 2 The triangulation of the (coarse) computational domain Ω^f_0 and of the vocal fold model M5 (dimensions in [m]). From the top of vocal folds was chosen point $A = [0.00628; -0.00057]$

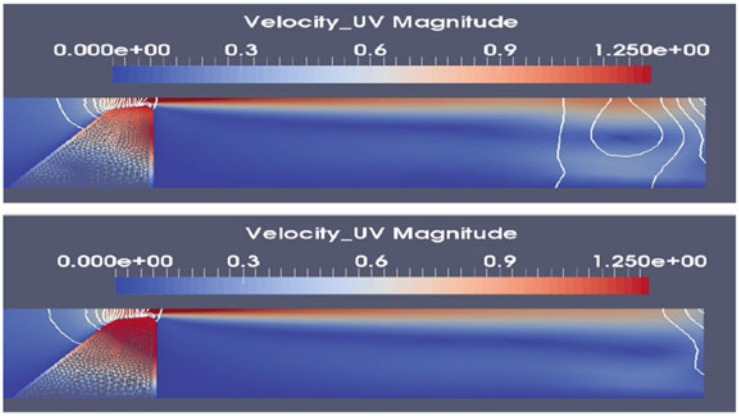


Fig. 3 The detail of the flow field at two time instants $t = 0.075$ and 0.45 s together with the pressure isolines. The order of the vocal fold displacements is 10^{-4} m

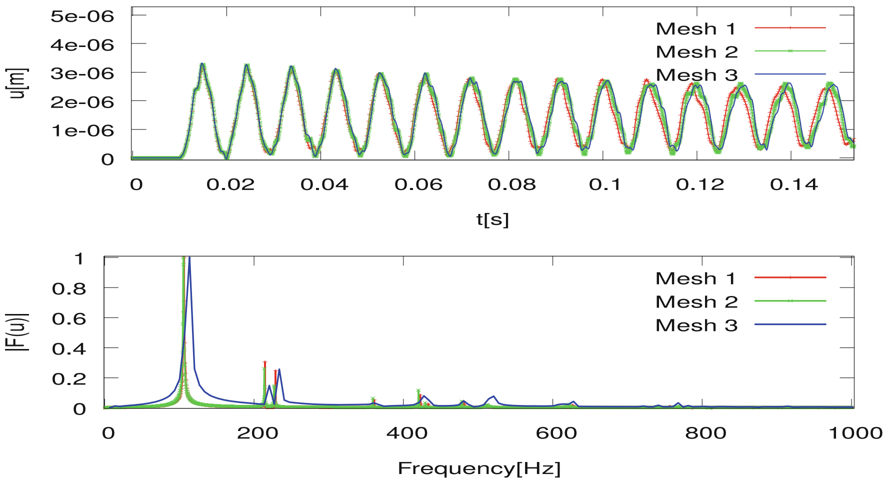


Fig. 4 The time signal for the displacement of point A in direction x and below its scaled Fourier transformation $|F(u)|$ without first 0.01 s. Mesh 1 is the most coarse one. Frequency spectrum was rescaled to have maximum at 1

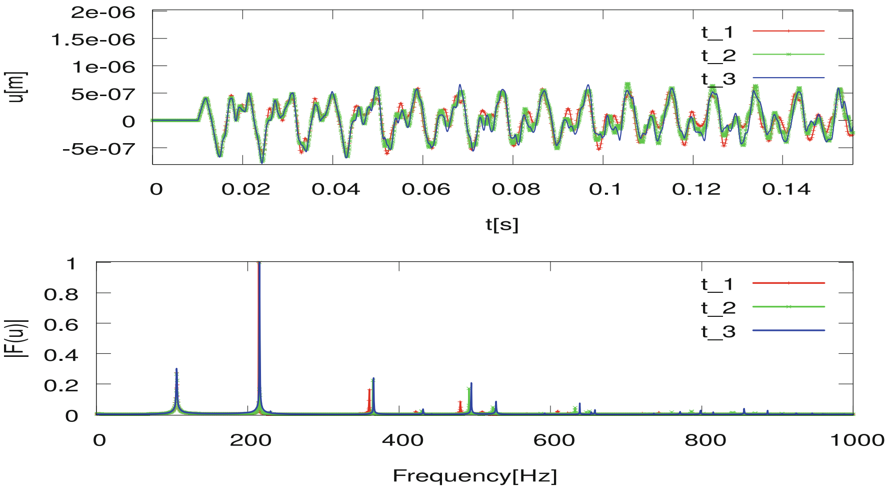


Fig. 5 The time signal of the displacement of point A in direction y and below its scaled Fourier transformation $|F(u)|$ with excluded first 0.01 s. Frequency spectrum was rescaled to have maximum at 1

The solution for three time steps $t_1 = 2 \cdot 10^{-4}$ s, $t_2 = 1 \cdot 10^{-4}$ s and $t_3 = 5 \cdot 10^{-5}$ s was analyzed on the coarse mesh. The results are shown in Fig. 5. It can be seen that agreement is good, the Fourier transformation confirms the excitation of mainly first two modes for all cases on time interval of 1 s. Similar result was achieved for example by [12].

5 Conclusion

This contribution described the mathematical model of the FSI problem and its numerical approximation. The simple test case was computed and the convergence of presented method in time and space was demonstrated. The results showed that this method is for given case sufficiently accurate if the appropriately fine grid and enough small time step are used.

Acknowledgements The financial support of this was partly provided by the *Czech Science Foundation* under the *Grant No. P101/11/0207*.

References

1. R.A. Adams, *Sobolev Spaces* (Academic, New York, 1975)
2. M. Braack, P.B. Mucha, Directional do-nothing condition for the Navier-Stokes equations. J. Comput. Math. **32**, 507–521 (2014)
3. R. Clark, E.H. Dowell, *A Modern Course in Aeroelasticity* (Springer, 2004). http://www.springer.com/us/book/9781402027116
4. A. Curnier, *Computational Methods in Solid Mechanics* (Springer, Dordrecht/Boston, 1994)
5. T.A. Davis, *Direct Methods for Sparse Linear Systems* (SIAM, Philadelphia, 2006)
6. V. Girault, P.A. Raviart, *Finite Element Methods for Navier-Stokes Equations* (Springer, Berlin/New York, 1986)
7. R.C. Scherer, D. Shinwari, K.J. De Witt, C. Zhang, B.R. Kucinschi, A.A. Afjeh, Intraglottal pressure profiles for a symmetric and oblique glottis with a divergence angle of 10 degrees. J. Acoust. Soc. Am. **109**, 1616–1630 (2001)
8. P. Sváček, J. Horáček, Numerical simulation of glottal flow in interaction with self oscillating vocal folds: comparison of finite element approximation with a simplified model. Commun. Comput. Phys. **12**, 789–806 (2012)
9. N. Takashi, T.J.R. Hughes, An arbitrary Lagrangian-Eulerian finite element method for interaction of fluid and a rigid body. Comput. Methods Appl. Mech. Eng. **95**, 115–138 (1992)
10. I.R. Titze, F. Alipour, *The Myoelastic Aerodynamic Theory of Phonation* (National Center for Voice and Speech, Denver/Iowa City, 2006)
11. J. Valášek, P. Sváček, J. Horáček, Numerical simulation of interaction of fluid flow and elastic structure modelling vocal fold. Appl. Mech. Mater. **821**, 693–700 (2016)
12. S. Zörner, M. Kaltenbacher, M. Döllinger, Investigation of prescribed movement in fluid-structure interaction simulation for the human phonation process. Comput. Fluids **86**, 133–140 (2013)

Part IX
Miscellaneous Topics

Comparison of Nonlocal Operators Utilizing Perturbation Analysis

Burak Aksoylu and Fatih Celiker

Abstract We present a comparative study of integral operators used in nonlocal problems. The size of nonlocality is determined by the parameter δ. The authors recently discovered a way to incorporate local boundary conditions into nonlocal problems. We construct two nonlocal operators which satisfy local homogeneous Neumann boundary conditions. We compare the bulk and boundary behaviors of these two to the operator that enforces nonlocal boundary conditions. We construct approximations to each operator using perturbation expansions in the form of Taylor polynomials by consistently keeping the size of expansion neighborhood equal to δ. In the bulk, we show that one of these two operators exhibits similar behavior with the operator that enforces nonlocal boundary conditions.

1 Introduction

The integral operators under consideration are used, for instance, in peridynamics (PD) [11] and nonlocal diffusion [5, 8]. PD is a nonlocal extension of continuum mechanics developed by Silling [11]. PD is based on nonlocal interactions. As a result, nonlocal boundary conditions (BC) are used. The authors recently discovered a way to incorporate local BC into nonlocal theories [3, 4, 7], in particular into PD.

We present a comparative study of operators used in nonlocal problems. We consider problems in 1D and choose the domain $\Omega = (-1, 1)$. We define the

B. Aksoylu (✉)
Department of Mathematics, TOBB University of Economics and Technology, Ankara, 06560, Turkey

Department of Mathematics, Wayne State University, 656 W. Kirby, Detroit, MI 48202, USA
e-mail: baksoylu@etu.edu.tr

F. Celiker
Department of Mathematics, Wayne State University, 656 W. Kirby, Detroit, MI 48202, USA
e-mail: celiker@wayne.edu

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_57

governing operator related to PD by

$$Au(x) := cu(x) - \int_{\Omega} C(x-y)u(y)dy, \quad x \in \Omega, \tag{1}$$

where $C, u \in L^2(\Omega)$ and $c = \int_{\Omega} C(y)dy$. The kernel function $C(x)$ is assumed to be even. An important first choice of $C(x)$ is the *canonical* kernel function $\chi_\delta(x)$ whose only role is the representation of the nonlocal neighborhood, called the *horizon*, by a characteristic function. Namely,

$$\chi_\delta(x) := \begin{cases} 1, \ |x| < \delta \\ 0, \ \text{otherwise.} \end{cases}$$

The size of nonlocality is determined by δ and we assume $\delta < 1$.

In $\mathbb{R}^d, d \geq 1$, we discovered that the PD governing operator (1) is a bounded function of the classical (Laplace) operator [7]. We generalized this theoretical result to bounded domains [3, 4]. The main idea of the generalization is as follows. Building on the theoretical result, we generalized the standard integral based convolution in (1) to an abstract convolution operator which is defined by a Hilbert (complete and orthonormal) basis. This basis is induced by the classical operator with prescribed local BC on bounded domains. The nonlocal operator becomes a function of the classical operator. By prescribing BC to the classical operator, we construct a gateway to incorporate local BC into nonlocal theories.

Through the use of local BC, we plan to solve important elasticity applications which require local BC such as contact, shear, and traction. In addition, we anticipate to eliminate the surface effects which are seen in PD due to employing nonlocal BC. Incorporation of local BC leads to a modification of the original PD governing operator in (1).

The operators $\mathscr{M}$ and $\mathscr{N}$ defined below employ the even part of u. For notational convenience, we denote the orthogonal projections that give the even and odd parts, respectively, of a function by $P_e, P_o : L^2(\Omega) \to L^2(\Omega)$, whose definitions are

$$P_e u(x) := \frac{u(x) + u(-x)}{2}, \quad P_o u(x) := \frac{u(x) - u(-x)}{2}. \tag{2}$$

In this paper, we present a comparative study of the following three operators. For $x \in \Omega$,

$$\mathscr{L}u(x) := cu(x) - \int_{\Omega} \chi_\delta(x-y)u(y)dy, \tag{3}$$

$$\mathscr{M}u(x) := cu(x) - \int_{\Omega} \hat{\chi}_\delta(x-y)P_e u(y)dy, \tag{4}$$

$$\mathscr{N}u(x) := cu(x) - \int_{\Omega} \hat{\chi}_\delta(|x-y|-1)P_e u(y)dy. \tag{5}$$

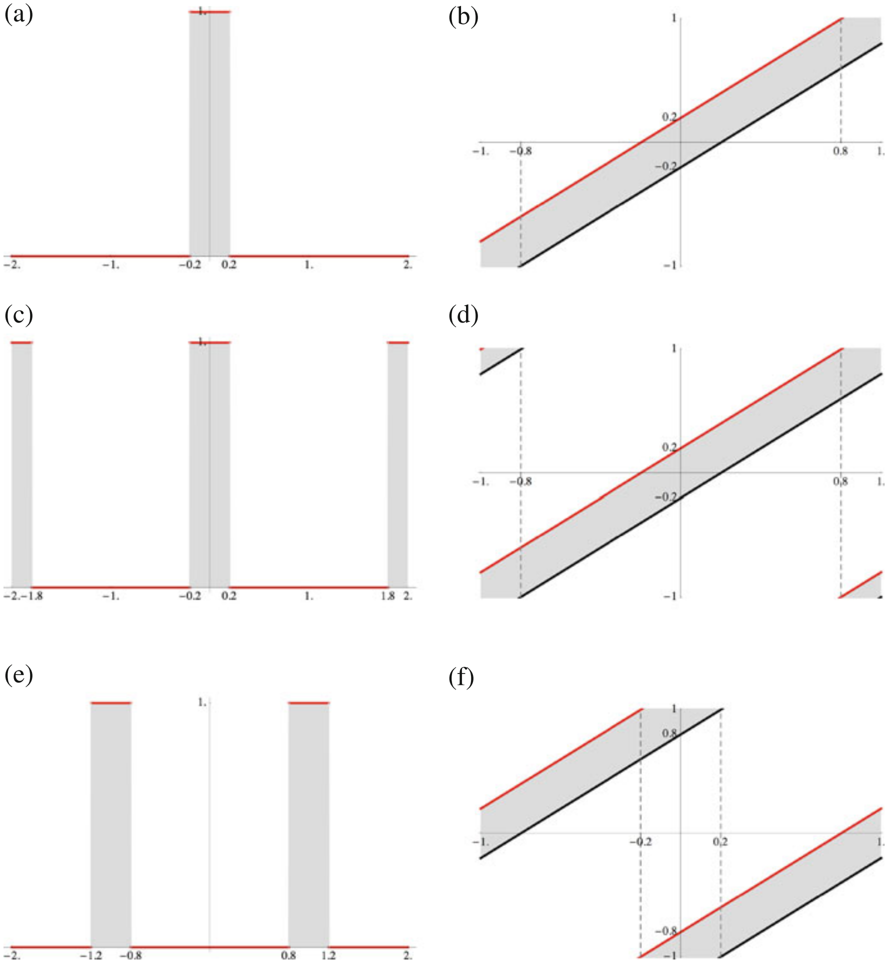


Fig. 1 Kernel functions are obtained by extensions from $\Omega = (-1, 1)$ to $\hat{\Omega} = (-2, 2)$ with $\delta = 0.2$ and their corresponding supports. (**a**) Kernel function $\chi_\delta(x)$. (**b**) Support of $\chi_\delta(x-y)$. (**c**) Kernel function $\hat{\chi}_\delta(x)$, periodic extension of $\chi_\delta(x)|_{x\in\Omega}$ to $\hat{\Omega}$. (**d**) Support of $\hat{\chi}_\delta(x-y)$. (**e**) Kernel function $\hat{\chi}_\delta(1-|x|)$. (**f**) Support of $\hat{\chi}_\delta(1-|x-y|)$

Here, we define the extended domain $\hat{\Omega} := (-2, 2)$ and denote the periodic extension of $\chi_\delta(x)|_{x\in\Omega}$ to $\hat{\Omega}$ by $\hat{\chi}_\delta(x)$; see Fig. 1. We construct approximations $\widetilde{\mathscr{L}}, \widetilde{\mathscr{M}}, \widetilde{\mathscr{N}}$ to each governing operator $\mathscr{L}, \mathscr{M}, \mathscr{N}$ using perturbation expansions. Similar expansions were used by the first author [1, 2] and in higher order gradient applications [6, 9, 10].

2 Operator Definitions

For $x \in (-2, 2)$, kernel functions in (3), (4), and (5) are defined as

$$\chi_\delta(x) := \begin{cases} 1, \; x \in (-\delta, \delta) \\ 0, \; \text{otherwise.} \end{cases}$$

$$\hat{\chi}_\delta(x) := \begin{cases} 1, \; x \in (-2, -2+\delta) \cup (-\delta, \delta) \cup (2-\delta, 2) \\ 0, \; \text{otherwise.} \end{cases}$$

$$\hat{\chi}_\delta(1-|x|) := \begin{cases} 1, \; x \in (-1-\delta, -1+\delta) \cup (1-\delta, 1+\delta) \\ 0, \; \text{otherwise.} \end{cases}$$

The corresponding convolution kernels are

$$\chi_\delta(y-x) := \begin{cases} 1, \; y \in (x-\delta, x+\delta) \\ 0, \; \text{otherwise.} \end{cases}$$

$$\hat{\chi}_\delta(y-x) := \begin{cases} 1, \; y \in (x-2, x-2+\delta) \cup (x-\delta, x+\delta) \cup (x+2-\delta, x+2) \\ 0, \; \text{otherwise.} \end{cases}$$

$$\hat{\chi}_\delta(1-|y-x|) := \begin{cases} 1, \; y \in (x-1-\delta, x-1+\delta) \cup (x+1-\delta, x+1+\delta) \\ 0, \; \text{otherwise.} \end{cases}$$

With a slight abuse of notation, for functions $\hat{\chi}_\delta(\cdot) : \hat{\Omega} \to \mathbb{R}$ and its bivariate version $\hat{\chi}_\delta(\cdot,\cdot) : \Omega \times \Omega \to \mathbb{R}$, we use the same notation; $\hat{\chi}_\delta(x-y) = \hat{\chi}_\delta(x,y)$. In Fig. 1, we depict the support of $\hat{\chi}_\delta(x,y)$.

For integration, we need to consider the following y-ranges

$$\begin{aligned} \Omega_{\mathscr{L}} &:= (-1,1) \cap \{(x-\delta, x+\delta)\}, \\ \Omega_{\mathscr{M}} &:= (-1,1) \cap \{(x-2, x-2+\delta) \cup (x-\delta, x+\delta) \cup (x+2-\delta, x+2)\}, \\ \Omega_{\mathscr{N}} &:= (-1,1) \cap \{(x-1-\delta, x-1+\delta) \cup (x+1-\delta, x+1+\delta)\}. \end{aligned}$$

2.1 Boundary Conditions

The classical operator satisfying homogeneous Neumann BC is given by

$$A_{\mathrm{N}} u = -\frac{4}{\pi^2} u'',$$

where $'$ denotes the weak derivative and $u \in H_0^2(\Omega)$. A_{N} has a purely discrete spectrum $\sigma(A_{\mathrm{N}})$ consisting of simple eigenvalues,

$$\sigma(A_{\mathrm{N}}) = \left\{k^2 : k \in \mathbb{N}\right\}.$$

A normalized eigenvector corresponding to the eigenvalue k^2 is given by

$$e_k^{\mathrm{N}}(x) := \begin{cases} \frac{1}{\sqrt{2}}, & k = 0 \\ \cos\left(\frac{k\pi}{2}(x+1)\right), & k \neq 0, k \in \mathbb{N} \end{cases}.$$

The sequence $\left(e_k^{\mathrm{N}}\right)_{k\in\mathbb{N}}$ is a Hilbert basis of $L^2(\Omega)$. Using this basis, we define the generalized convolution operator on Ω for C, $u \in L^2(\Omega)$ [3, 4] as follows

$$\mathscr{C} *_{\mathrm{N}} u(x) := \sum_{k\in\mathbb{N}} \langle e_k^{\mathrm{N}} | C\rangle \, \langle e_k^{\mathrm{N}} | u\rangle \, e_k^{\mathrm{N}}(x), \tag{6}$$

where $\langle\cdot|\cdot\rangle$ denotes the inner product in $L^2(\Omega)$.

We want to obtain an integral representation for (6). For this, we need several ingredients. Let $\widehat{C}(x)$, $x \in (-2, 2)$ denote periodic extension of the kernel function $C(x)$, $x \in (-1, 1)$. Since $C(x)$ is even, so is $\widehat{C}(x)$. Then, $\widehat{C}(x) = \widehat{C}(|x|)$. The integral representation of $\mathscr{C} *_{\mathrm{N}}$ is based on the following decomposition of $\widehat{C}(|x|)$ based on the "half-wave symmetry."

$$\begin{aligned} \widehat{C}(|x|) &= \frac{\widehat{C}(|x|) + \widehat{C}(1-|x|)}{2} + \frac{\widehat{C}(|x|) - \widehat{C}(1-|x|)}{2}, \\ &=: \widehat{C}_1(x) + \widehat{C}_2(x). \end{aligned}$$

Then, the integral representation of $\mathscr{C} *_{\mathrm{N}}$ in (6) takes the following form [3, 4]

$$\mathscr{C} *_{\mathrm{N}} u(x) = \int_\Omega \widehat{C}(|x-y|-1) P_e u(y)dy + \gamma_{N,C} \int_\Omega u(y)dy, \tag{7}$$

where $\gamma_{N,C} := -\frac{\sqrt{2}-1}{2\sqrt{2}} \int_\Omega C_1(y)dy + \frac{\sqrt{2}+1}{2\sqrt{2}} \int_\Omega C_2(y)dy$. Hence,

$$\frac{d}{dx}\mathscr{C} *_{\mathrm{N}} u(x) = \frac{d}{dx}\int_\Omega \widehat{C}(|x-y|-1) P_e u(y)dy.$$

Observe that only the convolution part survives after differentiation. This allows us to induce several governing integral operators that satisfy homogeneous Neumann BC. As a result, we can obtain the operator $\mathscr{N}u$ in (5) with general kernel function $\widehat{C}(|x-y|-1)$

$$\mathscr{N}u(x) := cu(x) - \int_{\Omega} \widehat{C}(|x-y|-1)P_e u(y)dy. \tag{8}$$

Note that

$$\widehat{C}(|x|-1) = \widehat{C}_1(x) - \widehat{C}_2(x). \tag{9}$$

Using the fact that $\widehat{C}(x)$ is 2-periodic and after some algebraic manipulation, we conclude that both $\widehat{C}_1(x)$ and $\widehat{C}_2(x)$ are 2-periodic. Due to (9), $\widehat{C}(|x|-1)$ is also 2-periodic. Consequently, $\widehat{C}(|x|-1)$ is an even and 2-periodic function, a crucial property that we will also use for constructing the other governing operator; see (13).

Remark 1 $\widehat{C}_1(x)$ and $\widehat{C}_2(x)$ have an additional property of half-wave symmetry. Namely, for every $x \in [0, 1/2]$,

$$C_1(x) = \frac{1}{2}\left[C(x) + C(1-x)\right] = \frac{1}{2}\left[C(|1-x|) + C(1-|1-x|)\right] = C_1(1-x),$$

$$C_2(x) = \frac{1}{2}\left[C(x) - C(1-x)\right] = \frac{1}{2}\left[C(1-|1-x|) - C(|1-x|)\right] = -C_2(1-x).$$

These identities have been used in obtaining the integral representation in (7).

Next, we want to show that the governing operator in (8) satisfies the homogeneous Neumann BC. We begin with rewriting $\mathscr{N}u(x)$ as follows

$$\mathscr{N}u(x) = cu(x) - \int_{\Omega} (1/2)\left(\widehat{C}(|x-y|-1) + \widehat{C}(|x+y|-1)\right)u(y)]dy.$$

For simplicity, assuming that C is sufficiently smooth and differentiating both sides, we obtain

$$\frac{d}{dx}\mathscr{N}u(x) = cu'(x) - \\ \int_{\Omega} (1/2)\left(\widehat{C}'(|x-y|-1)\frac{|x-y|}{x-y} + \widehat{C}'(|x+y|-1)\frac{|x+y|}{x+y}\right)u(y)]dy. \tag{10}$$

The case of non-smooth C can be handled by splitting the integral into parts where C is piecewise smooth. Here, $u'(x)$ denotes the initial velocity, and hence, we always assume that it satisfies homogeneous Neumann BC, i.e., $u'(-1) = u'(1) = 0$ because initial values automatically satisfy the given BC. Furthermore, since $\widehat{C}(y)$

is even, $\widehat{C}'(y)$ is odd. Evaluating (10) at $x = -1$ gives

$$\begin{aligned}\frac{d}{dx}\mathscr{N}u(-1) &= cu'(-1) - \int_{\Omega}(1/2)\left(\widehat{C}'(y)(-1) + \widehat{C}'(-y)(-1)\right)u(y)dy\\ &= cu'(-1) - \int_{\Omega}(1/2)\left(\widehat{C}'(y)(-1) - \widehat{C}'(y)(-1)\right)u(y)dy\\ &= 0. \end{aligned} \tag{11}$$

Similarly, at $x = 1$, we have

$$\begin{aligned}\frac{d}{dx}\mathscr{N}u(1) &= cu'(1) - \int_{\Omega}(1/2)\left(\widehat{C}'(-y)(+1) + \widehat{C}'(y)(+1)\right)u(y)dy\\ &= cu'(1) - \int_{\Omega}(1/2)\left(-\widehat{C}'(y)(+1) + \widehat{C}'(y)(+1)\right)u(y)dy\\ &= 0. \end{aligned} \tag{12}$$

2.2 An Alternative Governing Operator

The main property we exploit in satisfying the BC is the evenness of the kernel function. Inspired by this fact, we can define a simpler alternative governing operator that satisfies homogeneous Neumann BC

$$\mathscr{M}u(x) := cu(x) - \int_{\Omega}\widehat{C}(x-y)P_e u(y)dy \tag{13}$$

$$= cu(x) - \int_{\Omega}(1/2)\left(\widehat{C}(x-y) + \widehat{C}(x+y)\right)u(y)]dy. \tag{14}$$

In a similar fashion to (11) and (12), one can easily show that (14) satisfies the BC.

3 Perturbation Expansions

We construct approximations $\widetilde{\mathscr{L}}, \widetilde{\mathscr{M}}, \widetilde{\mathscr{N}}$ to $\mathscr{L}, \mathscr{M}, \mathscr{N}$ using perturbation expansions in the form of Taylor polynomials by consistently keeping the size of the expansion neighborhood equal to δ in each case. This leads to Taylor polynomial of $u(y)$ at different y locations such as $y = x, -x, x-1, x+1$. That way, approximations of $u(y)$ all have error $\mathscr{O}(\delta^3)$, which means that we maintain consistent error among approximate operators.

For each operator, we have 3 intervals, on which the Taylor polynomials are guaranteed to have the size of the expansion neighborhood equal to δ in the corresponding y-range. We list the y-ranges, depict them in Fig. 1 as shaded regions. Then, we utilize a Taylor polynomial which defines the approximate integrand $\tilde{f}_{\cdot,\cdot}(x,y)$. Eventually, we calculate the approximate operator for the corresponding interval.

We easily see that $c = \int_{\Omega} \chi_{\delta}(y)dy = 2\delta$. For convenience, we prefer to use $\widetilde{\mathscr{L}}u(x) - 2\delta u(x)$, $\widetilde{\mathscr{M}}u(x) - 2\delta u(x)$, and $\widetilde{\mathscr{N}}u(x) - 2\delta u(x)$. The calculations for approximations $\widetilde{\mathscr{L}}, \widetilde{\mathscr{M}}, \widetilde{\mathscr{N}}$ are given in a systematic way. We also report values at transition points for each approximate operator.

3.1 Operator $\widetilde{\mathscr{L}}$

The integrand is $f_{\mathscr{L}}(x,y) = -u(y)$. We have 3 intervals, left, center, and right denoted by ℓ, $\mathtt{c}$, and $\mathtt{r}$, respectively. $I_{\mathscr{L},\ell} := (-1,-1+\delta)$, $I_{\mathscr{L},\mathtt{c}} := (-1+\delta, 1-\delta)$, and $I_{\mathscr{L},\mathtt{r}} := (1-\delta, 1)$.

3.1.1 Operator$\widetilde{\mathscr{L}}_{\ell}$, $x \in I_{\mathscr{L},\ell} = (-1,-1+\delta)$

$$y \in R_{\mathscr{L},\ell} = (-1, x+\delta),\ \ y - x \in (-x-1, \delta) \subset (-\delta,\delta),\ \ |y-x| < \delta$$

$$\begin{aligned}
R_{\mathscr{L},\ell} : u(y) &= u(x) + (y-x)u'(x) + \frac{(y-x)^2}{2}u''(x) + \mathscr{O}(\delta^3),\\
\tilde{f}_{\mathscr{L},\ell}(x,y) &= -u(x) - (y-x)u'(x) - \frac{(y-x)^2}{2}u''(x),\\
\widetilde{\mathscr{L}}_{\ell}u(x) - 2\delta u(x) &= \int\limits_{R_{\mathscr{L},\ell}} \tilde{f}_{\mathscr{L},\ell}(x,y)dy\\
&= [-x-1-\delta]u(x) + \frac{1}{2}\left[x-(-1-\delta)\right]\left[x-(-1+\delta)\right]u'(x)\\
&\quad + \frac{-1}{6}\left[x-(-1-\delta)\right]\left[x^2-(\delta-2)x+1-\delta+\delta^2\right]u''(x).
\end{aligned}$$

3.1.2 Operator$\widetilde{\mathscr{L}}_{\mathtt{c}}$, $x \in I_{\mathscr{L},\mathtt{c}} = (-1+\delta, 1-\delta)$

$$y \in R_{\mathscr{L},\mathtt{c}} = (x-\delta, x+\delta),\ \ y-x \in (-\delta,\delta),\ \ |y-x| < \delta$$

$$\tilde{f}_{\mathscr{L},\mathtt{c}}(x,y) = \tilde{f}_{\mathscr{L},\ell}(x,y),$$

$$\begin{aligned}\widetilde{\mathscr{L}}_{\mathrm{c}}u(x) - 2\delta u(x) &= \int\limits_{R_{\mathscr{L},\mathrm{c}}} \tilde{f}_{\mathscr{L},\mathrm{c}}(x,y)dy \\ &= \frac{-\delta^3}{3}u''(x) - 2\delta u(x).\end{aligned}$$

3.1.3 Operator$\widetilde{\mathscr{L}}_{\mathtt{r}}$, $x \in I_{\mathscr{L},\mathtt{r}} = (1-\delta, 1)$

$$y \in R_{\mathscr{L},\mathrm{r}} = (x-\delta, 1),\ \ y-x \in (-\delta, -x+1) \subset (-\delta, \delta),\ \ |y-x| < \delta$$

$$\begin{aligned}\tilde{f}_{\mathscr{L},\mathrm{r}}(x,y) &= \tilde{f}_{\mathscr{L},\ell}(x,y), \\ \widetilde{\mathscr{L}}_{\mathrm{r}}u(x) - 2\delta u(x) &= \int\limits_{R_{\mathscr{L},\mathrm{r}}} \tilde{f}_{\mathscr{L},\mathrm{r}}(x,y)dy \\ &= [x-1-\delta]u(x) + \frac{-1}{2}\left[x-(1+\delta)\right]\left[x-(1-\delta)\right]u'(x) \\ &\quad + \frac{1}{6}\left[x-(1+\delta)\right]\left[x^2+(\delta-2)x+1-\delta+\delta^2\right]u''(x).\end{aligned}$$

3.1.4 Values of$\widetilde{\mathscr{L}}$ at Boundary and Transition Points

$$\begin{aligned}\widetilde{\mathscr{L}}_{\ell}u(-1) &= \frac{-\delta^3}{6}u''(-1) + \frac{-\delta^2}{2}u'(-1) + \delta u(-1) \\ \widetilde{\mathscr{L}}_{\ell}u(-1+\delta) = \widetilde{\mathscr{L}}_{\mathrm{c}}u(-1+\delta) &= \frac{-\delta^3}{3}u''(-1+\delta) \\ \widetilde{\mathscr{L}}_{\mathrm{c}}u(1-\delta) = \widetilde{\mathscr{L}}_{\mathrm{r}}u(1-\delta) &= \frac{-\delta^3}{3}u''(1-\delta) \\ \widetilde{\mathscr{L}}_{\mathrm{r}}u(1) &= \frac{-\delta^3}{6}u''(1) + \frac{\delta^2}{2}u'(1) + \delta u(1).\end{aligned}$$

3.2 *Operator* $\widetilde{\mathscr{M}}$

The integrand is $f_{\mathscr{M}}(x,y) = -P_e u(y)$. Similar to the $\widetilde{\mathscr{L}}$ case, we have 3 intervals: $I_{\mathscr{M},\ell} := (-1, -1+\delta)$, $I_{\mathscr{M},\mathrm{c}} := (-1+\delta, 1-\delta)$, and $I_{\mathscr{M},\mathrm{r}} := (1-\delta, 1)$.

3.2.1 Operator $\widetilde{\mathscr{M}}_{\ell}$, $x \in I_{\mathscr{M},\ell} = (-1, -1+\delta)$

$$\begin{aligned}
&y \in R_{\mathscr{M},\ell} = (x+2-\delta, 1), \ y+x \in (2x+2-\delta, x+1) \subset (-\delta,\delta), \ |y+x| < \delta \\
&y \in R_{\mathscr{M},\ell-\mathrm{c}} = (-1, x+\delta), \ y-x \in (-x-1, \delta) \subset (-\delta,\delta), \qquad\qquad |y-x| < \delta.
\end{aligned}$$

$$\begin{aligned}
R_{\mathscr{M},\ell} : u(y) &= u(-x) + (y+x)u'(-x) + \frac{(y+x)^2}{2}u''(-x) + \mathscr{O}(\delta^3), \\
R_{\mathscr{M},\ell-\mathrm{c}} : u(y) &= u(x) + (y-x)u'(x) + \frac{(y-x)^2}{2}u''(x) + \mathscr{O}(\delta^3),
\end{aligned}$$

$$\begin{aligned}
\tilde{f}_{\mathscr{M},\ell}(x,y) &= -P_e u(x) + (y+x)P_o u'(x) - \frac{(y+x)^2}{2}P_e u''(x), \\
\tilde{f}_{\mathscr{M},\ell-\mathrm{c}}(x,y) &= -P_e u(x) - (y-x)P_o u'(x) - \frac{(y-x)^2}{2}P_e u''(x).
\end{aligned}$$

$$\begin{aligned}
\widetilde{\mathscr{M}}_{\ell}u(x) - 2\delta u(x) &= \int\limits_{R_{\mathscr{M},\ell}} \tilde{f}_{\mathscr{M},\ell}(x,y)dy + \int\limits_{R_{\mathscr{M},\ell-\mathrm{c}}} \tilde{f}_{\mathscr{M},\ell-\mathrm{c}}(x,y)dy \\
&= q_{\ell}(x)P_e u''(x) - (x+1-\delta)^2 P_o u'(x) - 2\delta P_e u(x). \qquad (15)
\end{aligned}$$

3.2.2 Operator $\widetilde{\mathscr{M}}_{\mathrm{c}}$, $x \in I_{\mathscr{M},\mathrm{c}} = (-1+\delta, 1-\delta)$

$$y \in R_{\mathscr{M},\mathrm{c}} = (x-\delta, x+\delta), \ y-x \in (-\delta,\delta), \ |y-x| < \delta$$

$$\begin{aligned}
\tilde{f}_{\mathscr{M},\mathrm{c}}(x,y) &= \tilde{f}_{\mathscr{M},\ell-\mathrm{c}}(x,y), \\
\widetilde{\mathscr{M}}_{\mathrm{c}}u(x) - 2\delta u(x) &= \int\limits_{R_{\mathscr{M},\mathrm{c}}} \tilde{f}_{\mathscr{M},\mathrm{c}}(x,y)dy \\
&= \frac{-\delta^3}{3}P_e u''(x) - 2\delta P_e u(x).
\end{aligned}$$

3.2.3 Operator $\widetilde{\mathscr{M}}_{\mathtt{r}}$, $x \in I_{\mathscr{M},\mathtt{r}} = (1-\delta, 1)$

$$\begin{array}{lll} y \in R_{\mathscr{M},\mathtt{r}} = (-1, x-2+\delta), & y+x \in (x-1, 2x-2+\delta) \subset (-\delta,\delta), & |y+x| < \delta \\ y \in R_{\mathscr{M},\mathtt{r-c}} = (x-\delta, 1), & y-x \in (-\delta, -x+1) \subset (-\delta,\delta), & |y-x| < \delta. \end{array}$$

$$\begin{aligned} \tilde{f}_{\mathscr{M},\mathtt{r}}(x,y) &= \tilde{f}_{\mathscr{M},\ell}(x,y), \\ \tilde{f}_{\mathscr{M},\mathtt{r-c}}(x,y) &= \tilde{f}_{\mathscr{M},\ell-\mathtt{c}}(x,y), \\ \widetilde{\mathscr{M}}_{\mathtt{r}}u(x) - 2\delta u(x) &= \int\limits_{R_{\mathscr{M},\mathtt{r}}} \tilde{f}_{\mathscr{M},\mathtt{r}}(x,y)dy + \int\limits_{R_{\mathscr{M},\mathtt{r-c}}} \tilde{f}_{\mathscr{M},\mathtt{r-c}}(x,y)dy \\ &= q_{\mathtt{r}}(x)P_e u''(x) + (x-1+\delta)^2 P_o u'(x) - 2\delta P_e u(x). \qquad (16) \end{aligned}$$

Remark 2 Expressions of the coefficients $q_\ell(x)$ and $q_{\mathtt{r}}(x)$ of $P_e u''(x)$ in (15) and (16), respectively, are quite involved. So, we prefer not to report them.

3.2.4 Values of $\widetilde{\mathscr{M}}$ at Boundary and Transition Points

$$\begin{aligned} \widetilde{\mathscr{M}}_\ell u(-1) - 2\delta u(-1) &= \frac{-\delta^3}{3} P_e u''(-1) - \delta^2 P_o u'(-1) - 2\delta P_e u(-1) \\ \widetilde{\mathscr{M}}_\ell u(-1+\delta) - 2\delta u(-1+\delta) &= \frac{-\delta^3}{3} P_e u''(-1+\delta) - 2\delta P_e u(-1+\delta) \\ \widetilde{\mathscr{M}}_{\mathtt{c}} u(-1+\delta) - 2\delta u(-1+\delta) &= \frac{-\delta^3}{3} P_e u''(-1+\delta) - 2\delta P_e u(-1+\delta) \\ \widetilde{\mathscr{M}}_{\mathtt{c}} u(1-\delta) - 2\delta u(1-\delta) &= \frac{-\delta^3}{3} P_e u''(1-\delta) - 2\delta P_e u(1-\delta) \\ \widetilde{\mathscr{M}}_{\mathtt{r}} u(1-\delta) - 2\delta u(1-\delta) &= \frac{-\delta^3}{3} P_e u''(1-\delta) - 2\delta P_e u(1-\delta) \\ \widetilde{\mathscr{M}}_{\mathtt{r}} u(1) - 2\delta u(1) &= \frac{-\delta^3}{3} P_e u''(1) + \delta^2 P_o u'(1) - 2\delta P_e u(1). \end{aligned}$$

3.3 *Operator* $\widetilde{\mathscr{N}}$

The integrand is $f_{\mathscr{N}}(x,y) = -P_e u(y)$. We have 3 intervals: $I_{\mathscr{N},\ell} := (-1,-\delta)$, $I_{\mathscr{N},\mathtt{c}} := (-\delta,\delta)$, and $I_{\mathscr{N},\mathtt{r}} := (\delta, 1)$.

3.3.1 Operator $\widetilde{\mathscr{N}}_{\ell}$, $x \in I_{\mathscr{N},\ell} = (-1, -\delta)$

$$y \in R_{\mathscr{N},\ell} = (x+1-\delta, x+1+\delta),\ \ y-(x+1) \in (-\delta,\delta),\ \ |y-(x+1)| < \delta.$$

$$\begin{aligned}
R_{\mathscr{N},\ell} : u(y) &= u(x+1) + [y-(x+1)]u'(x+1) \\
&\quad + \frac{[y-(x+1)]^2}{2}u''(x+1) + \mathscr{O}(\delta^3), \\
\tilde{f}_{\mathscr{N},\ell}(x,y) &= -P_e u(x+1) \\
&\quad -[y-(x+1)]P_o u'(x+1) - \frac{[y-(x+1)]^2}{2}P_e u''(x+1), \\
\widetilde{\mathscr{N}}_{\ell} u(x) - 2\delta u(x) &= \int\limits_{R_{\mathscr{N},\ell}} \tilde{f}_{\mathscr{N},\ell}(x,y)dy \\
&= \frac{-\delta^3}{3}P_e u''(x+1) - 2\delta P_e u(x+1).
\end{aligned}$$

3.3.2 Operator $\widetilde{\mathscr{N}}_{\mathtt{c}}$, $x \in I_{\mathscr{N},\mathtt{c}} = (-\delta, \delta)$

$$y \in R_{\mathscr{N},\ell-\mathtt{c}} = (x+1-\delta, 1), y-(x+1) \in (-\delta,-x) \subset (-\delta,\delta), |y-(x+1)| < \delta$$
$$y \in R_{\mathscr{N},\mathtt{r}-\mathtt{c}} = (-1, x-1+\delta), y-(x-1) \in (-x,\delta) \subset (-\delta,\delta), |y-(x-1)| < \delta.$$

$$\begin{aligned}
R_{\mathscr{N},\ell-\mathtt{c}} : u(y) &= u(x+1) + [y-(x+1)]u'(x+1) + \frac{[y-(x+1)]^2}{2}u''(x+1) + \mathscr{O}(\delta^3) \\
R_{\mathscr{N},\mathtt{r}-\mathtt{c}} : u(y) &= u(x-1) + [y-(x-1)]u'(x-1) + \frac{[y-(x-1)]^2}{2}u''(x-1) + \mathscr{O}(\delta^3)
\end{aligned}$$

$$\begin{aligned}
\tilde{f}_{\mathscr{N},\ell-\mathtt{c}}(x,y) &= \tilde{f}_{\mathscr{N},\ell}(x,y), \\
\tilde{f}_{\mathscr{N},\mathtt{r}-\mathtt{c}}(x,y) &= -P_e u(x-1) - [y-(x-1)]P_o u'(x-1) - \frac{[y-(x-1)]^2}{2}P_e u''(x-1).
\end{aligned}$$

$$\widetilde{\mathscr{N}}_{\mathtt{c}}u(x) - 2\delta u(x) = \int\limits_{R_{\mathscr{N},\ell-\mathtt{c}}} \tilde{f}_{\mathscr{N},\ell-\mathtt{c}}(x,y)dy + \int\limits_{R_{\mathscr{N},\mathtt{r}-\mathtt{c}}} \tilde{f}_{\mathscr{N},\mathtt{r}-\mathtt{c}}(x,y)dy$$

$$= \frac{1}{6}(x^3-\delta^3)P_e u''(x+1) - \frac{1}{6}(x^3+\delta^3)P_e u''(x-1) +$$

$$\frac{1}{2}(x^2-\delta^2)[-P_o u'(x+1) + P_o u'(x-1)] + (x-\delta)P_e u(x+1) - (x+\delta)P_e u(x-1).$$

3.3.3 Operator $\widetilde{\mathscr{N}}_{\mathtt{r}}$, $x \in I_{\mathscr{N},\mathtt{r}} = (\delta, 1)$

$$y \in R_{\mathscr{N},\mathtt{r}} = (x-1-\delta, x-1+\delta),\ \ y-(x-1) \in (-\delta,\delta),\ \ |y-(x-1)| < \delta.$$

$$\begin{aligned} \tilde{f}_{\mathscr{N},\mathtt{r}}(x,y) &= \tilde{f}_{\mathscr{N},\mathtt{r}-\mathtt{c}}(x,y), \\ \widetilde{\mathscr{N}}_{\mathtt{r}}u(x) - 2\delta u(x) &= \int\limits_{R_{\mathscr{N},\mathtt{r}}} \tilde{f}_{\mathscr{N},\mathtt{r}}(x,y)dy \\ &= \frac{-\delta^3}{3}P_e u''(x-1) - 2\delta P_e u(x-1). \end{aligned}$$

3.3.4 Values of $\widetilde{\mathscr{N}}$ at Boundary and Transition Points

$$\begin{aligned} \widetilde{\mathscr{N}}_{\ell}u(-1) - 2\delta u(-1) &= \frac{-\delta^3}{3}P_e u''(0) - 2\delta P_e u(0) \\ \widetilde{\mathscr{N}}_{\ell}u(-\delta) - 2\delta u(-\delta) &= \frac{-\delta^3}{3}P_e u''(1-\delta) - 2\delta P_e u(1-\delta) \\ \widetilde{\mathscr{N}}_{\mathtt{c}}u(-\delta) - 2\delta u(-\delta) &= \frac{-\delta^3}{3}P_e u''(1-\delta) - 2\delta P_e u(1-\delta) \\ \widetilde{\mathscr{N}}_{\mathtt{c}}u(\delta) - 2\delta u(\delta) &= \frac{-\delta^3}{3}P_e u''(\delta-1) - 2\delta P_e u(\delta-1) \\ \widetilde{\mathscr{N}}_{\mathtt{r}}u(\delta) - 2\delta u(\delta) &= \frac{-\delta^3}{3}P_e u''(\delta-1) - 2\delta P_e u(\delta-1) \\ \widetilde{\mathscr{N}}_{\mathtt{r}}u(1) - 2\delta u(1) &= \frac{-\delta^3}{3}P_e u''(0) - 2\delta P_e u(0). \end{aligned}$$

4 Comparison of Operators

4.1 Comparison in the Bulk

The interval $(-1+\delta, 1-\delta)$ is usually referred as the bulk of the domain. The behavior in the bulk is considered to be the main behavior of the operator especially when $\delta \ll 1$. That is why, it is important to find out the operator behavior in the bulk. By construction, the notion of bulk is slightly different for the $\mathscr{N}$ operator. The intervals $(-1,-\delta)$ and $(\delta, 1)$ will be referred as bulk in the case of $\mathscr{N}$. We list the bulk behavior of each operator:

$$\widetilde{\mathscr{L}}_{\mathrm{c}} u(x) = \frac{-\delta^3}{3} u''(x), \quad x \in (-1+\delta, 1-\delta), \tag{17}$$

$$\widetilde{\mathscr{M}}_{\mathrm{c}} u(x) - 2\delta u(x) = \frac{-\delta^3}{3} P_e u''(x) - 2\delta P_e u(x), \quad x \in (-1+\delta, 1-\delta), \tag{18}$$

$$\widetilde{\mathscr{N}}_{\ell} u(x) - 2\delta u(x) = \frac{-\delta^3}{3} P_e u''(x+1) - 2\delta P_e u(x+1), \quad x \in (-1,-\delta), \tag{19}$$

$$\widetilde{\mathscr{N}}_{\mathrm{r}} u(x) - 2\delta u(x) = \frac{-\delta^3}{3} P_e u''(x-1) - 2\delta P_e u(x-1), \quad x \in (\delta, 1). \tag{20}$$

We start comparing the operators with $\widetilde{\mathscr{L}}_{\mathrm{c}}$ and $\widetilde{\mathscr{M}}_{\mathrm{c}}$. Then, by substituting $u = P_e u$ in (18) and using $P_e^2 = P_e$, we arrive at

$$\widetilde{\mathscr{M}}_{\mathrm{c}} P_e u(x) = \frac{-\delta^3}{3} P_e u''(x). \tag{21}$$

In order to match (17) with (21), we also substitute $u = P_e u$ and we get

$$\widetilde{\mathscr{L}}_{\mathrm{c}} P_e u(x) = \frac{-\delta^3}{3} P_e u''(x).$$

Then, we conclude that the action of $\widetilde{\mathscr{L}}_{\mathrm{c}}$ and $\widetilde{\mathscr{M}}_{\mathrm{c}}$ agree in the bulk when restricted to the even component of $u(x)$.

In order to compare $\widetilde{\mathscr{N}}_{\ell}$ and $\widetilde{\mathscr{N}}_{\mathrm{r}}$ with $\widetilde{\mathscr{L}}_{\mathrm{c}}$, we substitute $u = P_e u$ in (19) and (20), which gives us the following results:

$$\widetilde{\mathscr{N}}_{\ell} P_e u(x) - 2\delta P_e u(x) = \frac{-\delta^3}{3} P_e u''(x+1) - 2\delta P_e u(x+1) \tag{22}$$

$$\widetilde{\mathscr{N}}_{\mathrm{r}} P_e u(x) - 2\delta P_e u(x) = \frac{-\delta^3}{3} P_e u''(x-1) - 2\delta P_e u(x-1). \tag{23}$$

In order to cancel the $2\delta P_e u(x)$ with $2\delta P_e u(x+1)$ and $2\delta P_e u(x-1)$ in (22) and (23), respectively, we need to make the following assumption:

$$u(x) = u(x-1) = u(x+1), \quad x \in (-1, -\delta) \cup (\delta, 1). \tag{24}$$

This property holds, for instance, when u is 1-periodic. Namely,

$$u(x) = u(x-1), \quad x \in \mathbb{R}. \tag{25}$$

We may conclude that the assumption (24) is triggered because of the half-wave symmetry property, noted in Remark 1, employed when constructing the integral operator $\mathscr{N}$. In summary, we conclude that $\widetilde{\mathscr{L}}_{\mathtt{c}}$ agrees with $\widetilde{\mathscr{N}}_{\ell}$ and $\widetilde{\mathscr{N}}_{\mathtt{r}}$ when restricted to the even component of $u(x)$ where $u(x)$ is 1-periodic.

4.2 Comparison of Higher Order Approximations in the Bulk

If we use a higher order Taylor approximation, for instance, for the $\widetilde{\mathscr{L}}_{\ell}$ operator $x \in I_{\mathscr{L},\ell} = (-1, -1+\delta)$, we get following expansion of $y \in R_{\mathscr{L},\ell} = (-1, x+\delta)$

$$u(y) = \left(I + (y-x)D + \cdots + \frac{(y-x)^{2n}}{(2n)!} D^{2n} \right) u(x) + \mathscr{O}(\delta^{2n+1}).$$

Then the error of the following operators is $\mathscr{O}(\delta^{2n+2})$.

$$\begin{aligned}
\widetilde{\mathscr{L}}_{\mathtt{c}} u(x) &= (-2)\left(\frac{\delta^3}{3!} D^2 + \frac{\delta^5}{5!} D^4 + \cdots + \frac{\delta^{2n+1}}{(2n+1)!} D^{2n} \right) u(x), \\
\widetilde{\mathscr{M}}_{\mathtt{c}} P_e u(x) &= (-2)\left(\frac{\delta^3}{3!} D^2 + \frac{\delta^5}{5!} D^4 + \cdots + \frac{\delta^{2n+1}}{(2n+1)!} D^{2n} \right) P_e u(x), \\
\widetilde{\mathscr{N}}_{\ell} P_e u(x) &= (-2)\left(\frac{\delta^3}{3!} D^2 + \frac{\delta^5}{5!} D^4 + \cdots + \frac{\delta^{2n+1}}{(2n+1)!} D^{2n} \right) P_e u(x+1), \\
\widetilde{\mathscr{N}}_{\mathtt{r}} P_e u(x) &= (-2)\left(\frac{\delta^3}{3!} D^2 + \frac{\delta^5}{5!} D^4 + \cdots + \frac{\delta^{2n+1}}{(2n+1)!} D^{2n} \right) P_e u(x-1).
\end{aligned}$$

Note that all these expressions on the right hand side can be written as a function of D^2. This is an indication that all the above approximate operators are functions of the Laplace; see the extended discussion in [7].

4.3 Comparison at the Boundary and Transition Points

We monitor where we can capture the factor $\frac{-\delta^3}{3}$ next to $u''(x)$ and $P_e u''(x)$ terms. We consider this as an indication that the bulk behavior is captured at that point. By using transition values computed in Sect. 3.1, first note that

$$\begin{aligned}\widetilde{\mathscr{L}}_\ell u(-1+\delta) &= \widetilde{\mathscr{L}}_{\mathtt{c}} u(-1+\delta) = \frac{-\delta^3}{3} u''(-1+\delta)\\ \widetilde{\mathscr{L}}_{\mathtt{r}} u(1-\delta) &= \widetilde{\mathscr{L}}_{\mathtt{c}} u(1-\delta) = \frac{-\delta^3}{3} u''(1-\delta).\end{aligned}$$

At transition points $x = -1+\delta, 1-\delta$, we conclude that we can define a continuous extension of $\widetilde{\mathscr{L}}$ from the pieces $\widetilde{\mathscr{L}}_\ell, \widetilde{\mathscr{L}}_{\mathtt{c}}$, and $\widetilde{\mathscr{L}}_{\mathtt{r}}$.

In order to monitor boundary and bulk behavior, we need to manipulate boundary and transition expressions of the $\widetilde{\mathscr{M}}$ given in Sect. 3.2 by $u = P_e u$. Then, using implications of (2), i.e., $P_e P_o = 0$ and $P_e^2 = P_e$, we obtain

$$\begin{aligned}\widetilde{\mathscr{M}}_\ell P_e u(-1) &= \frac{-\delta^3}{3} P_e u''(-1)\\ \widetilde{\mathscr{M}}_\ell P_e u(-1+\delta) &= \frac{-\delta^3}{3} P_e u''(-1+\delta)\\ \widetilde{\mathscr{M}}_{\mathtt{c}} P_e u(-1+\delta) &= \frac{-\delta^3}{3} P_e u''(-1+\delta)\\ \widetilde{\mathscr{M}}_{\mathtt{r}} P_e u(1-\delta) &= \frac{-\delta^3}{3} P_e u''(1-\delta)\\ \widetilde{\mathscr{M}}_{\mathtt{r}} P_e u(1) &= \frac{-\delta^3}{3} P_e u''(1).\end{aligned}$$

Similar to the $\widetilde{\mathscr{L}}$ case, we can define a continuous extension of $\widetilde{\mathscr{M}} P_e$ at transition points from the pieces $\widetilde{\mathscr{M}}_\ell P_e, \widetilde{\mathscr{M}}_{\mathtt{c}} P_e$, and $\widetilde{\mathscr{M}}_{\mathtt{r}} P_e$.

In order to monitor boundary and bulk behavior of $\widetilde{\mathscr{N}}$, we manipulate boundary and transition expressions given in Sect. 3.3 by $u = P_e u$. Then, by assuming (25), we obtain

$$\begin{aligned}\widetilde{\mathscr{N}}_\ell P_e u(-1) &= \frac{-\delta^3}{3} P_e u''(-1)\\ \widetilde{\mathscr{N}}_\ell P_e u(-\delta) = \widetilde{\mathscr{N}}_{\mathtt{c}} P_e u(-\delta) &= \frac{-\delta^3}{3} P_e u''(-\delta)\\ \widetilde{\mathscr{N}}_{\mathtt{c}} P_e u(\delta) = \widetilde{\mathscr{N}}_{\mathtt{r}} P_e u(\delta) &= \frac{-\delta^3}{3} P_e u''(\delta)\\ \widetilde{\mathscr{N}}_{\mathtt{r}} P_e u(1) &= \frac{-\delta^3}{3} P_e u''(1).\end{aligned}$$

We can also define a continuous extension of $\widetilde{\mathscr{N}} P_e$ at transition points from the pieces $\widetilde{\mathscr{N}}_\ell P_e$, $\widetilde{\mathscr{N}}_{\mathtt{c}} P_e$, and $\widetilde{\mathscr{N}}_{\mathtt{r}} P_e$.

Note that values of $\widetilde{\mathscr{M}} P_e$ at boundary and bulk points exhibit the bulk behavior of $\widetilde{\mathscr{L}}$. In addition, under assumption (25), values of $\widetilde{\mathscr{N}}_{\mathtt{r}} P_e$ and $\widetilde{\mathscr{N}}_\ell P_e$ at boundary points also exhibit the bulk behavior of $\widetilde{\mathscr{L}}$. These might be indications that the surface effect issue observed in PD can be eliminated if $\mathscr{M}$ and $\mathscr{N}$ are used as governing operators. This is a future research avenue.

5 Conclusion

The important property we seek is to obtain $-\delta^3/3$ as the coefficient of the term with the second derivative. We identify this as the bulk behavior. Both $\widetilde{\mathscr{M}} P_e$ and $\widetilde{\mathscr{N}} P_e$ exhibit the same bulk behavior as $\widetilde{\mathscr{L}} P_e$. Furthermore, the bulk behavior is also observed at all boundary and transition points for $\widetilde{\mathscr{M}} P_e$ and $\widetilde{\mathscr{N}} P_e$. The comparison of $\widetilde{\mathscr{N}}$ to $\widetilde{\mathscr{L}}$ and $\widetilde{\mathscr{M}}$ requires the assumption of (25). Due to this restriction, we conclude that $\widetilde{\mathscr{L}}$ agrees with $\widetilde{\mathscr{M}}$ more than it does with $\widetilde{\mathscr{N}}$.

In the expansion of $\widetilde{\mathscr{M}} P_e$, the coefficients of $P_e u''(x)$ are all equal to $-\delta^3/3$ at transition points as well as at boundary points. This can be interpreted as the best possible agreement with the Laplace operator. Such an agreement may indicate that the surface effects observed in PD can be eliminated especially when $\mathscr{M}$ is used as governing operator. For future research, by eliminating the assumptions $u = P_e u$ and (25), we plan to construct governing operators that agree with $\mathscr{L}$ in the bulk.

Acknowledgements Burak Aksoylu was supported in part by National Science Foundation DMS 1016190 grant, European Commission Marie Curie Career Integration Grant 293978, and Scientific and Technological Research Council of Turkey (TÜBİTAK) TBAG 112T240 and MAG 112M891 grants. Fatih Celiker's sabbatical visit was supported in part by the TÜBİTAK 2221 Fellowship for Scientist on Sabbatical Leave Program. He also would like to thank Orsan Kilicer of Middle East Technical University for his careful reading of the paper.

Fatih Celiker was supported in part by the National Science Foundation DMS 1115280 grant.

References

1. B. Aksoylu, M.L. Parks, Variational theory and domain decomposition for nonlocal problems. Appl. Math. Comput. **217**, 6498–6515 (2011)
2. B. Aksoylu, Z. Unlu, Conditioning analysis of nonlocal integral operators in fractional Sobolev spaces. SIAM J. Numer. Anal. **52**(2), 653–677 (2014)
3. B. Aksoylu, H.R. Beyer, F. Celiker, Application and implementation of incorporating local boundary conditions into nonlocal problems (Submitted)
4. B. Aksoylu, H.R. Beyer, F. Celiker, Theoretical foundations of incorporating local boundary conditions into nonlocal problems (Submitted)
5. F. Andreu-Vaillo, J.M. Mazon, J.D. Rossi, J. Toledo-Melero, *Nonlocal Diffusion Problems*. Mathematical Surveys and Monographs, vol. 165 (American Mathematical Society, Providence/Real Socied Matematica Espanola, Madrid, 2010)

6. M. Arndt, M. Griebel, Derivation of higher order gradient continuum models from atomistic models for crystalline solids. Multiscale Model. Simul. **4**(2), 531–562 (2005)
7. H.R. Beyer, B. Aksoylu, F. Celiker, On a class of nonlocal wave equations from applications. J. Math. Phys. **57**(6) (2016). doi:org/10.1063/1.4953252, http://scitation.aip.org/content/aip/journal/jmp/57/6/10.1063/1.4953252
8. Q. Du, M. Gunzburger, R.B. Lehoucq, K. Zhou, Analysis and approximation of nonlocal diffusion problems with volume constraints. SIAM Rev. **54**, 667–696 (2012)
9. E. Emmrich, O. Weckner, On the well-posedness of the linear peridynamic model and its convergence towards the Navier equation of linear elasticity. Commun. Math. Sci. **5**(4), 851–864 (2007)
10. P. Seleson, M.L. Parks, M. Gunzburger, R.B. Lehoucq, *Peridynamics as an upscaling of molecular dynamics*. Multiscale Model. Simul. **8**, 204–227 (2009)
11. S. Silling, Reformulation of elasticity theory for discontinuities and long-range forces. J. Mech. Phys. Solids **48**, 175–209 (2000)

Pricing of Basket Options Using Dimension Reduction and Adaptive Finite Differences in Space, and Discontinuous Galerkin in Time

Lina von Sydow, Paria Ghafari, Erik Lehto, and Mats Wångersjö

Abstract European basket options are priced by solving the multi-dimensional Black–Scholes–Merton equation. Standard numerical methods to solve these problems often suffer from the “curse of dimensionality”. We tackle this by using a dimension reduction technique based on a principal component analysis with an asymptotic expansion. Adaptive finite differences are used for the spatial discretization. In time we employ a discontinuous Galerkin scheme. The efficiency of our proposed method to solve a five-dimensional problem is demonstrated through numerical experiments and compared with a Monte-Carlo method.

1 Introduction

Pricing of options is something that’s going on daily in banks and financial institutes. For many options there exist no analytical solution to the pricing problem and fast and accurate numerical methods are of utmost importance.

We consider a Black–Scholes–Merton market [1, 8] with one risk free asset with price process $B(t)$ and d risky assets with processes $\mathbf{S}(t) = (S_1(t) \cdots S_d(t))$ given by the following dynamics

$$\begin{aligned} dB(t) &= rB(t)dt, \\ dS_1(t) &= \alpha_1 S_1(t)dt + \sigma_1 S_1(t)dW_1(t), \\ &\vdots \\ dS_d(t) &= \alpha_d S_d(t)dt + \sigma_d S_d(t)dW_d(t), \end{aligned}$$

L. von Sydow (✉) • P. Ghafari • M. Wångersjö
Department of Information Technology, Uppsala University, Uppsala, Sweden
e-mail: lina@it.uu.se; paria.ghafari@gmail.com; mats.wangersjo@gmail.com

E. Lehto
Department of Information Technology, Uppsala University, Uppsala, Sweden

Department of Mathematics, KTH Royal Institute of Technology, Stockholm, Sweden
e-mail: elehto@kth.se

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_58

where $r \in \mathbb{R}$ is the short rate of interest, and W_i are correlated Wiener processes with $\langle dW_i, dW_j \rangle = \rho_{ij} dt$. Finally $\alpha_i, \sigma_i \in \mathbb{R}$ are the local mean of return and volatility of S_i respectively. An option issued on $\mathbf{S}$ that at time of maturity T pays $\varPhi(\mathbf{S})$ can be priced from

$$u = e^{-r(T-t)} \mathbb{E}^Q \left[\varPhi(\mathbf{S}(T)) \right], \tag{1}$$

where $\mathbb{E}^Q [\cdot]$ denotes the expected value under the risk-neutral measure Q. A standard way to price multi-dimensional problems is to use a Monte-Carlo method, simulating paths of $S_i(t)$ combined with (1). This approach is known to converge very slowly.

In [1] and [8] it was independently shown that the price u of an option issued on the risky asset can be obtained from solving the Black–Scholes–Merton equation

$$\begin{aligned} \tfrac{\partial u}{\partial t} + \textstyle\sum_{i=1}^{d} r s_i \tfrac{\partial u}{\partial s_i} + \tfrac{1}{2} \textstyle\sum_{i,j=1}^{d} \sigma_i \sigma_j \rho_{ij} s_i s_j \tfrac{\partial^2 u}{\partial s_i \partial s_j} - ru = 0, \\ u(T, \mathbf{S}) = \varPhi(\mathbf{S}). \end{aligned} \tag{2}$$

We consider European basket options that at time of maturity pays

$$\varPhi(\mathbf{S}) = \max(\sum_{i=1}^{d} \mu_i s_i - K, 0), \tag{3}$$

where μ_i determines the fraction of asset S_i in the basket and K is the so called strike price of the option. A standard discretization of (2) leads to the "curse of dimensionality" – the number of degrees of freedom of the discretized problem grows exponentially in the number of dimensions d. We will therefore introduce a dimension reduction techique based on a principal component analysis (PCA) and asymptotic expansions.

The outline of the paper is as follows: In Sect. 2, the dimension reduction technique is presented. The discretization in space and time is discussed in Sects. 3 and 4 respectively. Section 5 is devoted to the presentation of numerical results and we give concluding remarks in Sect. 6.

2 Dimension Reduction

We follow [2, 3, 12] and make the following change of variables

$$\mathbf{x} = \mathbf{Q}^T ln \mathbf{S} + \bar{\mathbf{b}} \tau, \tag{4}$$

where $\tau = T - t$, $b_i = \sum_{j=1}^{d} q_{ji} (r - \frac{\sigma_j^2}{2})$ and $\mathbf{Q}$ is the eigenvector-matrix of the covariance matrix $\boldsymbol{\Sigma}$ with elements $\boldsymbol{\Sigma}_{i,j} = \sigma_i \sigma_j \rho_{ij}$. Implementing the changes to (2) and (3) gives

$$\begin{aligned} \tfrac{\partial u}{\partial \tau} - \tfrac{1}{2}\textstyle\sum_{i=1}^{d} \lambda_i \tfrac{\partial^2 u}{\partial x_i^2} + ru = 0 & \\ u(0,\mathbf{x}) = \max\left(\textstyle\sum_{i=1}^{d} \mu_i e^{\sum_{j=1}^{d} q_{ji}x_j} - K, 0\right), & \end{aligned} \tag{5}$$

where λ_i are the eigenvalues of $\boldsymbol{\Sigma}$. Here $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_d|$.

The dimension reduction process is completed by an asymptotic expansion where we approximate each of the non-principal dimensions by a linear asymptotic expansion. Following [2, 3, 12] the asymptotic expansion of the solution is given by

$$u = u^{(1)} + \sum_{i=2}^{d} \lambda_i \left.\frac{\partial u}{\partial \lambda_i}\right|_{\bar{\lambda}=\bar{\lambda}^{(1)}} + \mathscr{O}(||\bar{\lambda} - \bar{\lambda}^{(1)}||^2) \tag{6}$$

where $u^{(1)}$ is the solution to the one-dimensional problem in the principal direction (corresponding to the largest eigenvalue), $\bar{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_d)$, and $\bar{\lambda}^{(1)} = (\lambda_1, 0, \ldots, 0)$. The derivatives in (6) can be approximated by a finite difference method

$$\left.\frac{\partial u}{\partial \lambda_i}\right|_{\bar{\lambda}=\bar{\lambda}^{(1)}} = \frac{u^{(1,i)} - u^{(1)}}{\lambda_i} + \mathscr{O}(\lambda_i^2) \tag{7}$$

where $u^{(1,i)}$ is the solution to the two-dimensional problem on the plane spanned by the principal axis x_1 and axis i corresponding the ith largest eigenvalue. Thus, the d-dimensional problem is broken down to one one-dimensional and $(d-1)$ two-dimensional problems. From (6) and (7) we see that if the eigenvalues λ_i, $i = 2, \ldots, d$ are small, the error introduced from the expansion is small.

3 Adaptive Finite Differences in Space

The PCA and asympotic expansions lead to the following PDEs to solve

$$\frac{\partial u}{\partial \tau} = \mathscr{L}u, \tag{8}$$

where the one- and two-dimensional spatial operators are defined respectively by

$$\mathscr{L}u = \frac{1}{2}\lambda_1 \frac{\partial^2 u}{\partial x_1^2} - ru, \tag{9a}$$

$$\mathscr{L}u = \frac{1}{2}\lambda_1 \frac{\partial^2 u}{\partial x_1^2} + \frac{1}{2}\lambda_i \frac{\partial^2 u}{\partial x_i^2} - ru, i = 2, \ldots, d. \tag{9b}$$

We apply a discretization in space using finite differences on a structured but possibly non-equidistant grid. The number of grid points in dimension i is N_i, $i = 1, \dots, d$. The second derivative in direction i is approximated as

$$\frac{\partial^2 u(x_{ik})}{\partial x_k^2} \approx a_{ik} u(x_{i(k+1)}) + b_{ik} u(x_{ik}) + c_{ik} u(x_{i(k-1)}), \tag{10}$$

where $a_{ik} = \frac{2}{h_{ik}(h_{i,k-1}+h_{ik})}$, $b_{ik} = \frac{-2}{h_{i,k-1}h_{ik}}$, $c_{ik} = \frac{2}{h_{i,k-1}(h_{i,k-1}+h_{ik})}$. The approximation (10) is second-order accurate in space if there is a smooth variation of the grid such that $h_{i,k-1} = h_{i,k}(1 + \mathcal{O}(h_{i,k}))$.

We will use the computational domain $x_i^0 - 5 \le x_i \le x_i^0 + 5$, $i = 1, \dots, d$, where $\mathbf{x}^0 = \mathbf{Q}^T ln\mathbf{S}^0 + \bar{\mathbf{b}}T$. At the boundaries of the computational domain we need to impose some boundary conditions. In the principal axis we use Dirichlet boundary conditions, while we approximate the solution across the boundaries in the other dimensions to be linear

$$u = 0,\ x_1 = x_1^0 - 5, \tag{11a}$$

$$u = \sum_{i=1}^{d} \mu_i e^{\sum_{j=1}^{n} q_{ij}(x_j - b_j\tau)} - Ke^{-r\tau},\ x_1 = x_1^0 + 5, \tag{11b}$$

$$\frac{\partial^2 u}{\partial x_i^2} = 0, \begin{cases} x_i = x_i^0 - 5, \\ x_i = x_i^0 + 5, \end{cases} i = 2, \dots, d. \tag{11c}$$

Next, we introduce spatial adaptivity for (8) with (9a), i.e. the 1D-problem in the principal axis. The outline of the adaptive algorithm is the same as in e.g.[9]:

1. Solve the PDE once using a coarse equidistant grid with N_1^c grid-points.
2. Create a new spatial grid aiming to fulfil the required accuracy.
3. Solve the PDE using the new adaptive grid with N_1 grid-points.

We will here only briefly discuss how to construct the adaptive grid, for a thorough explanation, see [5, 6, 9, 10, 16] . Assume that for the computed solution u_h using space-step h, it holds that $u_h = u + h^2 c(x) + \mathcal{O}(h^3)$. Using the second order accuracy also in the local discretization error ψ_h we get $\psi_h = h^2\eta(x) + \mathcal{O}(h^3)$. Omitting higher order terms we obtain after some algebraic manipulations $\psi_h = (\delta_{2h} - \delta_h)/3$ where $\delta_h = A_h u_h$. Estimating $\psi_{\bar{h}}$ using space-steps $\bar{h}$ and $2\bar{h}$ gives $\eta(x) = \psi_{\bar{h}}(x)/\bar{h}^2(x)$ and we get $|\psi_h(x)| = \left|h^2(x) \cdot \psi_{\bar{h}}(x)/\bar{h}^2(x)\right|$. In order to control the local discretization error and keep $|\psi_h(x)| \le \epsilon$ for some $\epsilon > 0$, we use $h(x) = \bar{h}(x)\left(\epsilon/|\psi_{\bar{h}}(x)|\right)^{1/2}$. Since the local discretization error varies over time and we want to have the same spatial grid for all $0 < \tau < T$, we will use $\psi_{\bar{h}}(x) = \max(|\psi_{\bar{h}}^{(\frac{T}{3})}(x)|, |\psi_{\bar{h}}^{(\frac{2T}{3})}(x)|, |\psi_{\bar{h}}^{(T)}(x)|)$.

For the two-dimensional problems (8), and (9b), we will in the principal axis use the adaptive grid computed. In the other axis we will use an equidistant grid using $\sim N_1/5$ grid-points. For a small number of grid-points like this, steps 1. and 2. above cost relatively much and the usage of equidistant points is preferable.

4 Discontinuous Galerkin in Time

The spatial discretization defined in Sect. 3 results in d systems of ordinary differential equations

$$\tfrac{du(\tau)}{d\tau} = Au(\tau) + f(\tau), \ \ 0 \le t \le T, \, u(0) = u_0 = \Phi, \tag{12}$$

where the vector $f(\tau)$ contains the boundary conditions. We will use a dG scheme to discretize (12), as in [4, 7, 11, 14, 15, 17, 18]. In [15] it was shown that for option pricing problems, dG is superior to other time-stepping schemes such as backwards differentiation formula of order 2, Crank-Nicolson and Rannacher time-stepping.

Divide $(0,T)$ into M subintervals $\{I_m = (\tau_{m-1}, \tau_m)\}_{m=1}^M$ of size $k = T/M$. Define $\mathscr{P}^r(I_m)$ as the space of polynomials of degree r or less on the interval I_m and $\mathbb{U} = \{U : U_m \in \mathscr{P}^r(I_m)\}$ to be the finite element space containing the piecewise polynomials. In the dG method the finite element solution U is continuous within each time interval I_m, but may be discontinuous at the nodes $\tau_1, \ldots, \tau_{M-1}$. We define the one-sided limits of a piecewise continuous function $u(\tau)$ as $u_m^+ := \lim_{s\to 0^+} u(\tau_m + s)$, $u_m^- := \lim_{s\to 0^+} u(\tau_m - s)$ and the jump across τ_m as $[u_m] := u_m^+ - u_m^-$.

The solution of (12) using a dG method of degree r (with order of accuracy $2r+1$) can be obtained by finding $U \in \mathbb{U}$ such that $\int_{I_m} (\dot{U} - AU) w(\tau)\, d\tau + [U_{m-1}] w(\tau_{m-1}) = \int_{I_m} f w(\tau)\, d\tau$, for $m = 1, \ldots, M$, all $w(\tau) \in \mathbb{U}$ and $U_0^- = u_0$.

Let $\{\varphi\}_{j=0}^r$ be a basis of the polynomial space $\mathscr{P}_r(-1,1)$ and let time shape functions on time interval I_m be given by $\varphi_j \circ F_m^{-1}$, where the mapping F_m defines a linear mapping $F_m : (-1,1) \to I_m$. Expanding U_m in $\mathscr{P}_r(I_m)$ $U_m = \sum_{j=0}^r u_{m,j} (\varphi_j \circ F_m^{-1})$ and using the basis $\{\varphi\}_{j=0}^r$ as test functions $w(t)$, we get after some algebraic manipulations

$$\left(\mathbf{C} \otimes \mathbf{I} - \frac{k}{2} \mathbf{G} \otimes \mathbf{A}\right) \mathbf{u}^m = \frac{k}{2} \mathbf{f}^1 + \mathbf{f}^2, \tag{13}$$

where $\otimes$ denotes the Kronecker product, $\mathbf{u}$ is the coefficient vector of U_m, $\mathbf{u}^m = \left(u_{m,0} \ldots u_{m,r}\right)^T$, and

$$\begin{aligned} f_{m,i}^1 &:= \textstyle\int_{-1}^1 (f \circ F_m) \varphi_i \, d\tau, & C_{ij} &:= \textstyle\int_{-1}^1 \varphi_j' \varphi_i \, d\tau + \varphi_j(-1)\varphi_i(-1), \\ f_{m,i}^2 &:= \varphi_i(-1) \textstyle\sum_{j=0}^{r_m} \varphi_j(1) u_{m-1,j}, & G_{ij} &= \textstyle\int_{-1}^1 \varphi_j \varphi_i \, d\tau. \end{aligned}$$

The time-integration of the Dirichlet boundary condition at $x_1^0 + 5$ requires the integration of (11b). This is accomplished using the composite trapezoidal rule with 100 intervals within each time-step.

Equation (13) forms a linear system of size $(r + 1)N$ for each time step. Using the temporal shape functions $\varphi_i(\tau) = (i + 1/2)^{1/2} L_i(\tau)$, with L_i denoting the i-th Legendre polynomial on $(-1, 1)$, this system decouples into $r + 1$ linear systems of size N, [14]. We get $\mathbf{G} = \mathbf{I}$ and

$$C_{ij} = \nu_{ij} \left(i + \frac{1}{2} \right)^{1/2} \left(j + \frac{1}{2} \right)^{1/2}, \nu_{ij} = \begin{cases} (-1)^{i+j} & \text{if } j < i \\ 1 & \text{otherwise} \end{cases}, i,j = 0, \ldots, r.$$

It can be shown that $\mathbf{C}$ is diagonalizable in $\mathbb{C}$ at least for $0 \leq r \leq 100$, [17]. Thus there exists a matrix $\mathbf{Q} \in \mathbb{C}^{(r+1)\times(r+1)}$ such that $\mathbf{Q}^{-1}\mathbf{C}\mathbf{Q} = \boldsymbol{\Delta} = \operatorname{diag}(\delta_0, \ldots, \delta_r)$. Multiplying (13) by $\mathbf{Q}^{-1} \otimes \mathbf{I}$ from the left gives a block-diagonal system that decouples into

$$\left(\delta_j \mathbf{M} - \frac{k}{2} \mathbf{A} \right) \mathbf{w}_j^m = \mathbf{g}_j, \quad j = 0, \ldots, r,$$

where $\mathbf{w}^m := (\mathbf{Q}^{-1} \otimes \mathbf{I})\mathbf{u}^m$ and $\mathbf{g} := (\mathbf{Q}^{-1} \otimes \mathbf{I}) \left(\frac{k}{2} \mathbf{f}^1 + \mathbf{f}^2 \right)$. Hence, in each time step we have to solve $r + 1$ linear systems of size N, which greatly reduces the time of computation and the usage of computer memory compared to solving the whole system (13) of size $(r + 1)N$. The linear systems of equations are solved using restarted GMRES(6), [13], with an incomplete LU-factorization as preconditioner [9].

5 Numerical Results

We use an example from [12] and consider the highly correlated basket option defined by the parameters in Table 1. However, we consider a call option (3) while a put option is considered in [12]. We have $\bar{\lambda}$ =

Table 1 Parameters for five-dimensional problem considered

r	T	K	Equity	$\bar{S}^0$	μ	$\bar{\sigma}$	ρ_{ij}
0.05	1	1	Deutsche Bank	1	0.381	0.518	1.00 0.79 0.82 0.91 0.84
			Hypo -Vereinsbank	1	0.065	0.648	0.79 1.00 0.73 0.80 0.76
			Commerzbank	1	0.057	0.623	0.82 0.73 1.00 0.77 0.72
			Allianz	1	0.270	0.570	0.91 0.80 0.77 1.00 0.90
			Münchner Rück	1	0.227	0.530	0.84 0.76 0.72 0.90 1.00

$(1.4089, 0.1124, 0.1006, 0.0388, 0.0213)$ for this problem. When we employ adaptivity we have used $N_1^c = 21$ in the coarse initial solution.

The implementation was made in MATLAB and run on an Apple MacBook Pro with 3.1 GHz Dual-Core Intel Core i7, Turbo Boost up to 3.4 GHz and 16 GB SDRAM.

We start by comparing the solution obtained using a Monte-Carlo method to price the option with the value obtained using our proposed method.

- Monte-Carlo method with 10^9 sampling paths: $\underline{0.22461}$.
- Our proposed method with $N_1 = 2594$ and 20 time-steps of dG with $r = 1$: $\underline{0.22443}$.

We see that the error introduced by truncating the asymptotic expansion is in the order of $\sim 10^{-4}$. We aim for a discretization that does not make the error in the final solution substantially larger.

First, we study the effect of using adaptivity in the principal axis x_1. In Fig. 1a we display the error in the solution as a function of the number of grid-points in the principal axis. The CPU-time as a function of error is presented in Fig. 1b. Note, that the somewhat erratic convergence behaviour is due to the fact that we study the point-wise error in $\mathbf{x}^0$ only. From Fig. 1a, b it is clear that the rate of convergence is close to the expected second-order for both the equidistant grids and adaptive grids. The error is smaller using adaptive grids using a certain number of grid-points and the CPU-time to reach a certain accuracy is considerably smaller.

Next, we study the error introduced by the discretization in time. In Fig. 2a we present the error as a function of number of adaptive grid-points in the principal axis using 6 and 20 time-steps respectively. It is clear that the error in the solution is not increased by using fewer amount of time-steps. Finally, in Fig. 2b we present the CPU-time it takes to compute the solution as a function of error. It is obvious that we gain by using the smaller number of time-steps. With adaptive grid-points

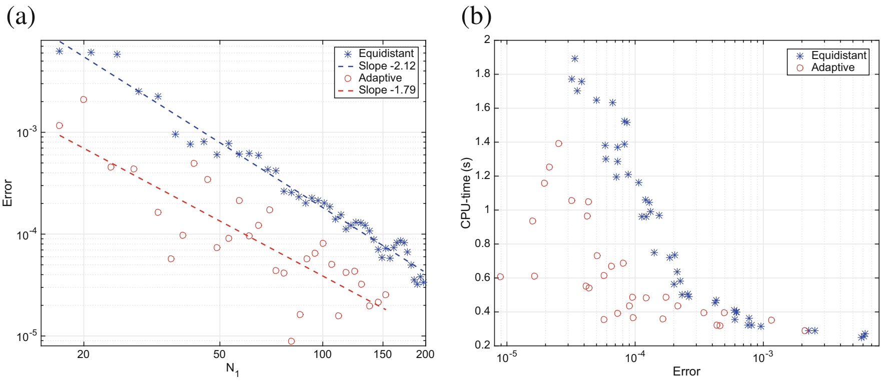


Fig. 1 Comparison of adaptive and equidistant grids. (**a**) Error as a function of N_1. (**b**) CPU-time as a function of error

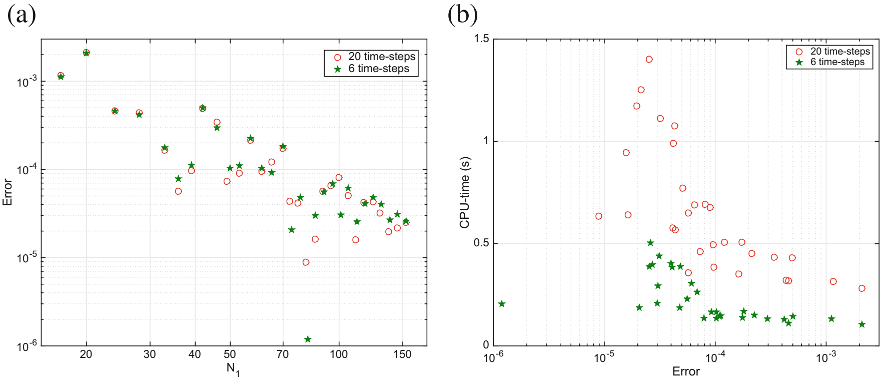


Fig. 2 Comparison of 6 and 20 time-steps with dG using $r = 1$. (**a**) Error as a function of N_1. (**b**) CPU-time as a function of error

in space and 6 time-steps of dG with $r = 1$ it takes less than 0.5 s to compute a solution that has an error $< 10^{-4}$ which is the accuracy that we aimed for. The time to compute a solution with the same accuracy level using a Monte-Carlo method is several minutes.

6 Conclusions

We consider the numerical solution of the multi-dimensional Black–Scholes–Merton equation to price basket options. A principal component analysis together with an asymptotic expansion is used to reduce the dimensionality of the underlying problem. The resulting PDEs are discretized in space with adaptive finite differences and a discontinuous Galerkin scheme in time. The efficiency of the proposed method is demonstrated for a five-dimensional basket option with highly correlated underlying assets.

References

1. F. Black, M. Scholes, The pricing of options and corporate liabilities. J. Polit. Econ. **81**, 637–654 (1973)
2. E. Ekedahl, E. Hansander, E. Lehto, *Dimension reduction for the Black-Scholes equation*, Tech. report, Department of Information Technology, Uppsala University (2007)
3. P. Ghafari, *Dimension reduction and adaptivity to price basket options*, Msc thesis, Uppsala University (2013). U.U.D.M. project report, 2013:3
4. E. Larsson, Option pricing using the discontinuous Galerkin method for time integration. Tech. report, Uppsala University (2013)

5. G. Linde, J. Persson, L. von Sydow, A highly accurate adaptive finite difference solver for the Black-Scholes equation. Int. J. Comput. Math. **86**, 2104–2121 (2009)
6. P. Lötstedt, J. Persson, L. von Sydow, J. Tysk, Space-time adaptive finite difference method for European multi-asset options. Comput. Math. Appl. **53**(8), 1159–1180 (2007)
7. A.-M. Matache, C. Schwab, T.P. Wihler, Fast numerical solution of parabolic integrodifferential equations with applications in finance. SIAM J. Sci. Comput. **27**(2), 369–393 (2005)
8. R.C. Merton, Theory of rational option pricing. Bell J. Econ. Manag. Sci. **4**, 141–183 (1973)
9. J. Persson, L. von Sydow, Pricing European multi-asset options using a space-time adaptive FD-method. Comput. Vis. Sci. **10**(4), 173–183 (2007)
10. J. Persson, L. von Sydow, Pricing American options using a space-time adaptive finite difference method. Math. Comput. Simul. **80**(9), 1922–1935 (2010)
11. T.V. Petersdorff, C. Schwab, Numerical solution of parabolic equations in high dimensions. ESAIM. Math. Model. Numer. Anal. **38**, 93–127 (2004)
12. C. Reisinger, G. Wittum, Efficient hierarchical approximation of high-dimensional option pricing problem. SIAM J. Sci. Comput. **29**, 440–458 (2007)
13. Y. Saad, M.H. Schultz, GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems. SIAM J. Sci. Stat. Comput. **7**(3), 856–869 (1986)
14. D. Schötzau, C. Schwab, Time discretization of parabolic problems by the *hp*-version of the discontinuous Galerkin finite element method. SIAM J. Numer. Anal. **38**, 837–875 (2000)
15. L. von Sydow, On discontinuous Galerkin for time integration in option pricing problems with adaptive finite differences in space, in *Numerical Analysis and Applied Mathematics: ICNAAM 2013, AIP Conference Proceedings*, vol. 1558 (American Institute of Physics (AIP), Melville, 2013), pp. 2373–2376
16. L. von Sydow, J. Toivanen, C. Zhang, Adaptive finite differences and IMEX time-stepping to price options under Bates model. Int. J. Comput. Math. **92**(12), 2515–2529 (2015)
17. T. Werder, K. Gerdes, D. Schötzau, C. Schwab, *hp*-discontinuous Galerkin time stepping for parabolic problems. Comput. Methods Appl. Mech. Eng. **190**, 6685–6708 (2001)
18. S. Zhao, G. Wei, A unified discontinuous Galerkin framework for time integration. Math. Methods Appl. Sci. **37**(7), 1042–1071 (2014)

On the Stability of a Weighted Finite Difference Scheme for Hyperbolic Equation with Integral Boundary Conditions

Jurij Novickij, Artūras Štikonas, and Agnė Skučaitė

Abstract We consider second order hyperbolic equation with nonlocal integral boundary conditions. We study the spectrum of the weighted difference operator for the formulated problem. Using the characteristic function we investigate the spectrum of the transition matrix of the three-layered finite difference scheme and obtain spectral stability conditions subject to boundary parameters γ_0, γ_1 and piecewise constant weight functions.

1 Introduction

In the theory of differential equations there often arise problems described by equations of mathematical physics with rather complicated nonclassical conditions modeling different life's processes. New applications are found in particle diffusion [1] and heat conduction [2]. Partial differential equations of the hyperbolic type with integral conditions often occur in problems related to fluid mechanics [3] (dynamics and elasticity), linear thermoelasticity [4], vibrations [5]. A survey on nonlocal boundary problems is presented in [6].

Consider the hyperbolic equation

$$\frac{\partial^2 u}{\partial t^2} - c^2 \frac{\partial^2 u}{\partial x^2} = f(x,t), \qquad (x,t) \in \Omega \times (0,T], \tag{1}$$

J. Novickij (✉) • A. Štikonas
Faculty of Mathematics and Informatics, Vilnius University, Naugarduko str. 24, LT-03225 Vilnius, Lithuania

Institute of Mathematics and Informatics, Vilnius University, Akademijos str. 4, LT-08663 Vilnius, Lithuania
e-mail: jurij.novickij@mif.vu.lt; arturas.stikonas@mif.vu.lt

A. Skučaitė
Institute of Mathematics and Informatics, Vilnius University, Akademijos str. 4, LT-08663 Vilnius, Lithuania
e-mail: agne.skucaite@mii.vu.lt

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_59

where $\Omega = (0, 1)$, with the classical initial conditions

$$u|_{t=0} = \phi(x), \ \left.\frac{\partial u}{\partial t}\right|_{t=0} = \psi(x), \quad x \in \overline{\Omega} := [0, 1], \tag{2}$$

and the additional nonlocal integral boundary conditions

$$u(0, t) = \gamma_0 \int_0^1 \beta^0(x)u(x, t)\, dx + v_l(t), \quad t \in [0, T], \tag{3}$$

$$u(1, t) = \gamma_1 \int_0^1 \beta^1(x)u(x, t)\, dx + v_r(t), \quad t \in [0, T], \tag{4}$$

where $f(x, t)$, $\phi(x)$, $\psi(x)$, $v_l(t)$, and $v_r(t)$ are given functions, γ_0 and γ_1 are given parameters, $\beta^0(x)$ and $\beta^1(x)$ are weight functions. Further we consider $c' = 1$ for simplicity. In this paper we focus on the weight functions of the type $\beta(x; a_0, a_1) = 1, a_0 \le x \le a_1$, and $\beta(x; a_0, a_1) = 0$, otherwise, where a_0, $a_1 \in [0, 1]$. We are interested in sufficiently smooth solutions of the nonlocal problem (1), (2), (3) and (4) (all the coefficients in (1), (2), (3) and (4) are smooth enough that the solution $U \in C^{4,4}$).

2 A Weighted Finite Difference Scheme

2.1 Notation

We introduce grids

$$\overline{\omega}^h := \left\{x_i : x_i = ih, i = \overline{0, n}\right\}; \ \overline{\omega}^\tau := \left\{t^j : t^j = j\tau, j = \overline{0, N}\right\};$$
$$\overline{\omega}^h_{1/2} := \left\{x_{i-1/2} = (x_{i-1} + x_i)/2, i = \overline{1, n}, x_{-1/2} = x_0, x_{n+1/2} = x_n\right\};$$
$$h = 1/n; \ \tau = T/N; \ h_{i+1/2} = x_{i+1/2} - x_{i-1/2}, i = \overline{0, n};$$
$$\omega^h := \{x_1, \ldots, x_{n-1}\}, \quad \tilde{\omega}^\tau := \left\{t^1, \ldots, t^N\right\}, \quad \omega^\tau := \left\{t^1, \ldots, t^{N-1}\right\};$$

where $n + 1$ and $N + 1$ are the numbers of grid points for x and t directions, accordingly, and $n, N \ge 2$.

We use the notation $U_i^j := U(x_i, t^j)$ for the function defined on the grid (or parts of the grid) $\overline{\omega}^h \times \overline{\omega}^\tau$. Instead of writing indices we denote $\breve{U}^j := U^{j-1}$ and $\widehat{U}^j := U^{j+1}$ on grids $\tilde{\omega}^\tau$ and $\omega^\tau \cup \{t^0\}$, respectively. Later in this paper we use the following notations $U^{(\sigma)} = \sigma_1 \breve{U} + (1 - \sigma_1 - \sigma_2)U + \sigma_2 \widehat{U}, \sigma_1, \sigma_2 \in \mathbb{R}$. We define a space grid operator

$$\delta_x^2 : \overline{\omega}^h \to \omega^h, \quad \left(\delta_x^2 U\right)_i := \frac{U_{i-1} - 2U_i + U_{i+1}}{h^2},$$

and the time grid operators

$$\overline{\partial}_t : \overline{\omega}^\tau \to \tilde{\omega}^\tau, \quad \overline{\partial}_t U := \frac{U - \breve{U}}{\tau}, \quad \overline{\partial}_t^2 : \overline{\omega}^\tau \to \omega^\tau, \quad \overline{\partial}_t^2 U := \frac{\breve{U} - 2U + \widehat{U}}{\tau^2}.$$

Let$\overline{H}$ and H be spaces of grid functions on $\overline{\omega}^h$ and ω^h, respectively. Similarly, let $\overline{H}_\tau$ and H_τ be spaces of grid functions on $\overline{\omega}^\tau$ and ω^τ. We also denote $H \times H_\tau$ as a space on $\omega^h \times \omega^\tau$. We define the inner products

$$[U, V] := \sum_{i=0}^{n} U_i V_i h_{i+1/2}, \quad U, V \in \overline{H}, \quad \textit{and} \quad (U, V) := \sum_{i=1}^{n-1} U_i V_i h, \quad U, V \in H.$$

Remark 1 For every function $U \in \overline{H}$ there exists a function $\mathring{U} \in H$, such that $\mathring{U} = U, \forall i = \overline{1, n-1}$.

2.2 Discrete Problem

Now we state a difference analogue of the differential problem (1), (2), (3) and (4). We define a weighted finite difference scheme (FDS) approximating the original differential equation (1):

$$\overline{\partial}_t^2 U - \delta_x^2 U^{(\sigma)} = F, \quad \left(x_i, t^j\right) \in \omega^h \times \omega^\tau, \tag{5}$$

where σ is a weight parameter of FDS. The initial conditions are approximated as follows:

$$U^0 = \Phi, \ \overline{\partial}_t U^1 = \Psi \quad x_i \in \overline{\omega}^h, \tag{6}$$

We rewrite the boundary conditions using the defined inner products:

$$U_0 = \gamma_0 [B^0, U] + V_l, \quad t^j \in \overline{\omega}^\tau, \quad U_n = \gamma_1 [B^1, U] + V_r, \quad t^j \in \overline{\omega}^\tau. \tag{7}$$

The functions B^0 and B^1 in the Eq. (7) correspond to the weight functions in Eqs. (3) and (4). Piecewise constant functions $\beta^0(x) = \beta(x; \xi_0^0, \xi_1^0)$ and $\beta^1(x) = \beta(x;\ \xi_0^1,\ \xi_1^1)$, can be replaced with the difference analogue $B_i^k = B_i^k(a_0, a_1) =$

1 if $a_0 \le x_i \le a_1$, and $B_i^k = 0$, otherwise. In the problem (5), (6) and (7) we approximate functions f, ϕ, ψ, v_l and v_r by grid functions $F \in H \times H_\tau$; $\Phi, \Psi \in \bar{H}$; and $V_l, V_r \in \bar{H}_\tau$.

Remark 2 We consider without loss of generality that functions $B_{\mathrm{k}i}(a_0, a_1)$ are defined on the uniform grid $\overline{\omega}^h$.

Remark 3 Both the boundary conditions (7) and the initial conditions (6) are defined at the points t^0 and t^1. At these points conditions are consistent. Properly choosing right hand side functions in (5), (6) and (7) one can obtain required approximation accuracy. For example, if $\Psi = \psi + 0.5\tau(\delta_x^2 U^0 + f^0)$ the differential problem (1), (2), (3) and (4) is approximated by (5), (6) and (7) with accuracy $\mathcal{O}(h^2)$.

2.3 The Three-Layer Finite Difference Scheme

Conditions (7) form a system of two linear equations for unknowns U_0 and U_n. We express these unknowns via inner points U_i, $i = \overline{1, n-1}$, and obtain

$$U_0 = \tilde{\gamma}_0(\tilde{B}^0, U) + \widetilde{V}_0, \quad U_n = \tilde{\gamma}_1(\tilde{B}^1, U) + \widetilde{V}_1, \tag{8}$$

where $\tilde{\gamma}_0 = \gamma_0 d^{-1}$, $\tilde{\gamma}_1 = \gamma_1 d^{-1}$ and

$$\tilde{B}^0 = \left(1 - \tfrac{h\gamma_1 B_n^1}{2}\right)B^0 + \tfrac{h\gamma_1 B_n^0}{2}B^1,\ \widetilde{V}_0 = d^{-1}\left(\tfrac{h\gamma_0}{2}B_n^0 V_r + \left(1 - \tfrac{h\gamma_1}{2}B_n^1\right)V_l\right),$$

$$\tilde{B}^1 = \tfrac{h\gamma_0 B_0^1}{2}B^0 + \left(1 - \tfrac{h\gamma_0 B_0^0}{2}\right)B^1,\ \widetilde{V}_1 = d^{-1}\left(\tfrac{h\gamma_1}{2}B_0^1 V_l + \left(1 - \tfrac{h\gamma_0}{2}B_0^0\right)V_r\right),$$

$$d = \frac{h^2\gamma_0\gamma_1}{4}\begin{vmatrix} B_0^0 & B_n^0 \\ B_0^1 & B_n^1 \end{vmatrix} - \frac{h}{2}(\gamma_0 B_0^0 + \gamma_1 B_n^1) + 1.$$

Problem (5), (8), according to [7, 8], can be transformed to the algebraic problem if $d \neq 0$. We have the following curves, displayed on Table 1, when the problem can not be transformed to the algebraic one.

By substituting expressions (8) into Eq. (5) for $i = 1$ and $i = n - 1$ we rewrite it in the form

$$\mathbf{A}\widehat{\mathbf{U}} + \mathbf{B}\mathbf{U} + \mathbf{C}\check{\mathbf{U}} = \tau^2\mathbf{F}, \tag{9}$$

$$\mathbf{A} = \mathbf{I} + \tau^2\sigma_1\boldsymbol{\Lambda},\ \mathbf{B} = -2\mathbf{I} + \tau^2(1 - \sigma_1 - \sigma_2)\boldsymbol{\Lambda},\ ,\mathbf{C} = \mathbf{I} + \tau^2\sigma_2\boldsymbol{\Lambda} \tag{10}$$

Table 1 Degeneration curves of the FDS with integral NBC with weights

Case	B_0^0	B_n^0	B_0^1	B_n^1	Degeneration curve	
1	0	0	0	0	Empty set	Ø
	0	0	1	0		
	0	1	0	0		
2	0	0	0	1	Horizontal line	$\gamma_1 = 2/h$
	0	0	1	1		
	0	1	0	1		
3	1	0	0	0	Vertical line	$\gamma_0 = 2/h$
	1	0	1	0		
	1	1	0	0		
4	1	0	0	1	Two lines	$\gamma_0 = 2/h$ or $\gamma_1 = 2/h$
	1	0	1	1		
	1	1	0	1		
5	0	1	1	0	Hyperbola	$\gamma_0\gamma_1 = 4/h^2$
	0	1	1	1		$h^2\gamma_0\gamma_1 + 2h\gamma_1 = 4$
	1	1	1	0		$h^2\gamma_0\gamma_1 + 2h\gamma_0 = 4$
6	1	1	1	1	Line	$\gamma_0 + \gamma_1 = 2/h$

where **A**, **B**, **C**, and

$$\boldsymbol{\Lambda} = \frac{1}{h^2}\begin{pmatrix} 2-ha_1 & -1-ha_2 & -a_3 & \dots & -ha_{n-3} & -ha_{n-2} & -ha_{n-1} \\ -1 & 2 & -1 & \dots & 0 & 0 & 0 \\ 0 & -1 & 2 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 2 & -1 & 0 \\ 0 & 0 & 0 & \dots & -1 & 2 & -1 \\ -hb_1 & -hb_2 & -hb_3 & \dots & -hb_{n-3} & -1-hb_{n-2} & 2-hb_{n-1} \end{pmatrix}, \tag{11}$$

are $(n-1)\times(n-1)$ matrices, **I** is the identity matrix, **0** is a zero matrix, $a_i = \tilde{\gamma}_0\tilde{B}_i^0$, $b_i = \tilde{\gamma}_1\tilde{B}_i^1$, $i = \overline{1, n-1}$. Finally, $\mathbf{F} = \big(\widetilde{F}_1, \dots, \widetilde{F}_{n-1}\big)^{\mathsf{T}}$, where $\widetilde{F}_i = F_i$, $i = \overline{2, n-2}$ and $\widetilde{F}_i = \widetilde{F}_i\big(F_i, V_l, V_r\big)$, $i = 1, n-1$.

2.4 The Two-Layer Finite Difference Scheme

We represent the three-layer scheme (9) as an equivalent two-layer scheme

$$\widehat{\mathbf{W}} = \mathbf{S}\mathbf{W} + \mathbf{G}, \tag{12}$$

using notations

$$\mathbf{W}=\begin{pmatrix}\mathbf{U}\\ \breve{\mathbf{U}}\end{pmatrix},\quad \mathbf{S}=\begin{pmatrix}-\mathbf{A}^{-1}\mathbf{B} & -\mathbf{A}^{-1}\mathbf{C}\\ \mathbf{I} & \mathbf{0}\end{pmatrix},\quad \mathbf{G}=\begin{pmatrix}\tau^2\mathbf{A}^{-1}\mathbf{F}\\ \mathbf{0}\end{pmatrix}. \tag{13}$$

One can study the stability conditions for the two-layer difference scheme (12) by analyzing the spectrum of the matrix $\mathbf{S}$. Note that the matrices $\mathbf{S}$ and $\boldsymbol{\Lambda}$ are nonsymmetric (matrix $\boldsymbol{\Lambda}$ is nonsymmetric except the classical case $\gamma_1 = 0$ and $\gamma_2 = 0$).

3 Spectrum Analysis

3.1 The Eigenstructure of the Matrix $\boldsymbol{\Lambda}$

One of the ways to investigate the stability of discrete problems is to study the eigenstructure of the transition matrix of finite difference scheme. So, we investigate an eigenvalue problem

$$\boldsymbol{\Lambda}\mathbf{U}=\lambda\mathbf{U},$$

for $(n-1)\times(n-1)$ matrix $\boldsymbol{\Lambda}$ which is in general equivalent to Sturm–Liouville Problem (SLP) for the difference operator with nonlocal boundary conditions

$$-\delta_x^2 U=\lambda U,\quad U\in\omega^h, \tag{14}$$

$$U_0=\gamma_0[B^0,U],\quad U_n=\gamma_1[B^1,U]. \tag{15}$$

Instead of investigating eigenvalues $\lambda\in\mathbb{C}_\lambda:=\mathbb{C}$ we use a bijection $\lambda=\lambda(q)$ from complex plane $\mathbb{C}_q$ to $\mathbb{C}_\lambda$:

$$\lambda=\frac{4}{h^2}\sin^2\frac{qh}{2},\quad q:=\alpha+\imath\beta \tag{16}$$

where $\mathbb{C}_q=\{q=\alpha\colon 0<\alpha<\pi/h\}\cup\{q=\imath\beta\colon\beta\geq 0\}\cup\{q=\pi/h+\imath\beta\colon\beta\geq 0\}$. The points $q=0$ and $q=\pi/h$ are the branch points of the map (16). So, every eigenvalue $\lambda_i=\lambda(q_i)$ conforms to q_i, $i=\overline{1,n-1}$ and vice versa.

Now we investigate the spectrum of matrix $\boldsymbol{\Lambda}$ in detail. The general solution of (14) in the case of $q\neq 0$, $q\neq\pi/h$ is $U=C_0\cos(qx)+C_1\sin(qx)$, $x\in\overline{\omega}^h$. By substituting it into (15) we have

$$\begin{aligned}\big(\gamma_0[B^0,\cos(qx)]-1\big)C_0+\gamma_0[B^0,\sin(qx)]C_1&=0,\\ \big(\gamma_1[B^1,\cos(qx)]-\cos q\big)C_0+\big(\gamma_1[B^1,\sin(qx)]-\sin q\big)C_1&=0.\end{aligned} \tag{17}$$

A nontrivial solutions of system (17) exist if its determinant is equal to zero:

$$\gamma_0\gamma_1 \begin{vmatrix} [B^0, \cos(qx)] & [B^0, \sin(qx)] \\ [B^1, \cos(qx)] & [B^1, \sin(qx)] \end{vmatrix} + \gamma_0 \begin{vmatrix} \cos q & \sin q \\ [B^0, \cos(qx)] & [B^0, \sin(qx)] \end{vmatrix} \\ - \gamma_1 [B^1, \sin(qx)] + \sin q = 0.$$

3.2 *The Structure of the Spectrum of the Matrix* S

In this section we investigate the relation between eigenvalues and eigenvectors of the two transition matrices $\boldsymbol{\Lambda}$ and $\mathbf{S}$. We denote $\lambda_k(\mathbf{A})$, $\lambda_k(\mathbf{B})$ and $\lambda_k(\mathbf{C})$ as the k-th eigenvalue of matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ accordingly. We denote that the matrices $\lambda_k(\mathbf{A})$, $\lambda_k(\mathbf{B})$ and $\lambda_k(\mathbf{C})$ commute (see [9]).

Let μ be the eigenvalue of the matrix $\mathbf{S}$. Consider the eigenvalue problem

$$\det(\mathbf{S} - \mu\mathbf{I}) = \det \begin{pmatrix} -\mathbf{A}^{-1}\mathbf{B} - \mu\mathbf{I} & -\mathbf{A}^{-1}\mathbf{C} \\ \mathbf{I} & -\mu\mathbf{I} \end{pmatrix} = \det \mathbf{A}^{-1} \det(\mathbf{A}\mu^2 + \mathbf{B}\mu + \mathbf{C}) = 0.$$

The above defined determinant is simplified to get a characteristic equation for the eigenvalues of the generalized nonlinear eigenvalue problem

$$(\mu^2\mathbf{A} + \mu\mathbf{B} + \mathbf{C})\mathbf{U} = 0, \quad \mathbf{U} \neq \mathbf{0}. \tag{18}$$

We note that the eigenvalues μ of the matrix $\mathbf{S}$ coincide with the eigenvalues of the generalized nonlinear eigenvalue problem (18). Let us clarify the relationship between the eigenvalues μ of the matrix $\mathbf{S}$ and the eigenvalues λ of the matrix $\boldsymbol{\Lambda}$. We denote $\mathbf{V}_k$ as an eigenvector of matrix $\boldsymbol{\Lambda}$. Substituting it into Eq. (18) we obtain

$$\big(\mu^2\mathbf{A} + \mu\mathbf{B} + \mathbf{C}\big)\mathbf{V}_k = \big(\mu^2\lambda_k(\mathbf{A}) + \mu\lambda_k(\mathbf{B}) + \lambda_k(\mathbf{C})\big)\mathbf{V}_k = 0. \tag{19}$$

So, eigenvalues of the matrix $\mathbf{S}$ satisfy the quadratic equation:

$$\mu^2\lambda_k(\mathbf{A}) + \mu\lambda_k(\mathbf{B}) + \lambda_k(\mathbf{C}) = 0, \quad k = \overline{1, N-1}. \tag{20}$$

Equation (20) determines the relation between eigenvalues μ_k^m and λ_k. The value of μ_k^m can be complex as well as real, depending on the parameters σ_1, σ_2, τ and eigenvalues λ_k.

According to the root criterion (see [10]) the roots of the second order polynomial $a\mu^2 + b\mu + c$ are in the closed unit disc of complex plane and those roots of magnitude 1 are simple if

$$|c|^2 + |\overline{a}b - \overline{b}c| \leq |a|^2, \quad |b| < 2|a|. \tag{21}$$

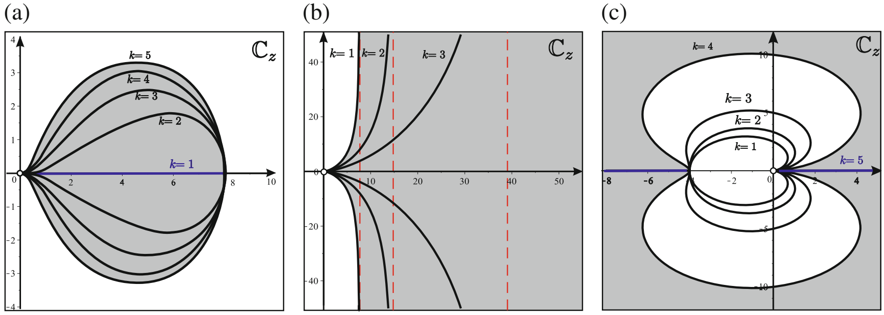


Fig. 1 Stability regions for different values of weights σ_1 and σ_2. (**a**) $\sigma_1 = 0.125k$, $k = \overline{1,5}$, $\sigma_1 + \sigma_2 = 0.25$. (**b**) $\sigma_1 = 0.48 - 0.05k$, $k = \overline{1,3}$, $\sigma_1 + \sigma_2 = 0.5$. (**c**) $\sigma_2 = 0.1k$, $k = \overline{1,5}$, $\sigma_1 + \sigma_2 = 1$

The quadratic polynomial coefficients for Eq. (20) are of the form: $a = 1 + \sigma_1 z$, $b = -(2 - (1 - \sigma_1 - \sigma_2)z)$, $c = 1 + \sigma_2 z$, $z = \tau^2\lambda$. Expressing z from (20) and substituting $\mu = e^{\imath\varphi}$, $\varphi \in (-\pi, +\pi]$ we obtain the formula for the boundary of the stability region:

$$z = z_\partial(\varphi) = \frac{2(1-\cos\varphi)\,(1-(\sigma_1+\sigma_2)(1-\cos\varphi)-(\sigma_1-\sigma_2)\imath\sin\varphi)}{(1-(\sigma_1+\sigma_2)(1-\cos\varphi))^2+(\sigma_1-\sigma_2)\sin^2\varphi}. \qquad (22)$$

In the work [11] we fully investigated the stability regions for the case of $\sigma_1 \neq \sigma_2$ and obtained a result, that the stability region exists only if $\sigma_1 \geq \sigma_2$. If a spectrum has complex eigenvalues, under the condition $\sigma = \sigma_1 = \sigma_2$, then FDS is unstable.

One can see the stability regions under various γ_1 and γ_2 parameters on the Fig. 1a–c. We notice, that Rez_∂ is even and Imz_∂ is odd functions, so the stability region is symmetric with respect to the real axis. The stability region is inside the boundary ∂S for $\sigma_1 + \sigma_2 < 1/2$ (see Fig. 1a), and outside the boundary ∂S for $\sigma_1 + \sigma_2 > 1/2$ (see Fig. 1c). The boundary points $z \in \partial S \setminus \{0\}$ belongs to the stability region (see Fig. 1a). In the case $\sigma_1 + \sigma_2 = 1/2$ (and $\sigma_1 \neq \sigma_2$) boundary ∂S divides complex plane into two unbounded parts (see Fig. 1b). The stability region is in the right-hand-side of the complex plane for $\sigma_1 > \sigma_2$.

Example 4 Let us take boundary conditions of the form $u(0,t) = 0$ and $u(1,t) = \gamma_1 \int_{1/4}^{3/4} u(x,t)\,dx$ (differential SLP was studied in [12]). The spectrum of formulated discrete problem (integrals approximated with trapezoid formula) was investigated in work [13]. The study is based on the investigation of characteristic curves on the part of a complex plane C_q, where $\lambda = 4/h^2 \sin^2(\pi qh/2h)$ (Fig. 2). The points of the spectrum belongs to a spectrum curves. These curves $\mathcal{N}_j$, $j = \overline{1,7}$ are shown in Fig. 2b, c. Every spectrum point moves along the spectrum curve while $\gamma \in (-\infty, +\infty)$. One can compare Fig. 2c with the stability regions shown in Fig. 1, keeping in mind relation $z = \tau^2\lambda$.

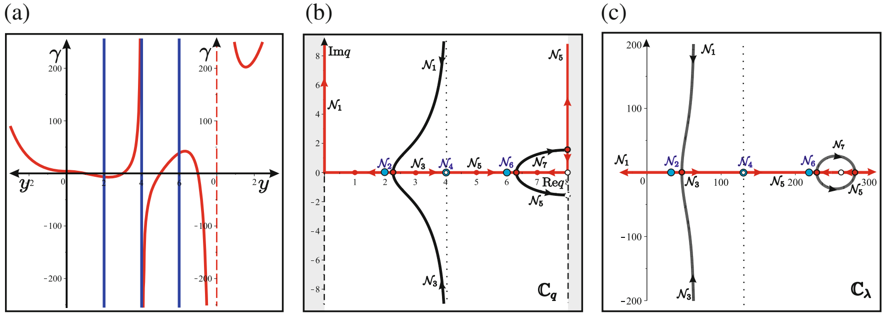


Fig. 2 Eigenspectrum of the numerical problem with one integral NBC. (**a**) Characteristic function. (**b**) Spectrum view on $\mathbb{C}_q$. (**c**) Spectrum view on $\mathbb{C}_\lambda$

The same situation is general for NBCs with not full integrals (see [13]). Except some the special cases complex eigenvalues exist.

Corollary 5 *FDS is unstable for sufficiently small $\tau \le \tau^*$ if the corresponding SLP has complex eigenvalues.*

Remark 6 If the corresponding SLP has complex eigenvalues then FDS can be stable for some $\tau > 0$ values' intervals only if we select special σ_1 and σ_2 values in the case $\sigma_1 > \sigma_2$, $\sigma_1 + \sigma_2 > 0.5$, $\gamma_0 = 0$, and $\gamma_1 < \gamma_*$. In the case $\gamma_0 \neq 0$ and $\gamma_1 \neq 0$ situation is more complex (see system (17)) and is under investigation.

Acknowledgements The research was partially supported by the Research Council of Lithuania (grant No. MIP-047/2014).

References

1. C. Mu, D. Liu, S. Zhou, Properties of positive solutions for a nonlocal reaction-diffusion equation with nonlocal nonlinear boundary condition. J. Korean Math. Soc. **6**(47), 1317–1328 (2010)
2. J.R. Cannon, The solution of the heat equation subject to the specification of energy. Q. Appl. Math. **21**, 155–160 (1963)
3. A.M. Nakhushev, An approximate method for solving boundary value problems for differential equations and its application to the dynamics of ground moisture and ground water. Differ. Equ. **18**(1), 72–81 (1982)
4. W. Day, *Heat Conduction Within Linear Thermoelasticity* (Springer, New York, 1985)
5. V.F. Volkodavov, V.E. Zhukov, Two problems for the string vibration equation with integral conditions and special matching conditions on the characteristic. Differ. Equ. **34**, 501–505 (1998)
6. A. Štikonas, A survey on stationary problems, Green's functions and spectrum of Sturm–Liouville problem with nonlocal boundary conditions. Nonlinear Anal. Model. Control **19**(3), 301–334 (2014)

7. J. Novickij, A. Štikonas, On the equivalence of discrete Sturm–Liouville problem with nonlocal boundary conditions to the algebraic eigenvalue problem. Proc. Lith. Math. Soc. Ser. A **56**, 66–71 (2015)
8. A. Štikonas, Investigation of characteristic curve for Sturm–Liouville problem with nonlocal boundary conditions on torus. Math. Model. Anal. **19**(4), 1–22 (2014)
9. J. Novickij, A. Štikonas, On the stability of a weighted finite difference scheme for wave equation with nonlocal boundary conditions. Nonlinear Anal. Model. Control **19**(3), 460–475 (2014)
10. A. Štikonas, The root condition for polynomial of the second order and a spectral stability of finite-difference schemes for Kuramoto–Tsuzuki equations. Math. Model. Anal. **3**, 214–226 (1998)
11. J. Novickij, A. Štikonas, On the stability of a finite difference scheme with two weights for wave equation with nonlocal conditions. Proc. Lith. Math. Soc. Ser. A **55**, 22–27 (2014)
12. R. Čiupaila, Ž. Jesevičūtė, M. Sapagovas, On the eigenvalue problem for one-dimensional differential operator with nonlocal integral condition. Nonlinear Anal. Model. Control **9**(2), 109–116 (2004)
13. A. Skučaitė, A. Štikonas, Spectrum curves for Sturm–Liouville problem with integral boundary condition. Math. Model. Anal. **20**(6), 802–818 (2015)

A Riemannian BFGS Method for Nonconvex Optimization Problems

Wen Huang, P.-A. Absil, and Kyle A. Gallivan

Abstract In this paper, a Riemannian BFGS method is defined for minimizing a smooth function on a Riemannian manifold endowed with a retraction and a vector transport. The method is based on a Riemannian generalization of a cautious update and a weak line search condition. It is shown that, the Riemannian BFGS method converges (i) globally to a stationary point without assuming that the objective function is convex and (ii) superlinearly to a nondegenerate minimizer. The weak line search condition removes completely the need to consider the differentiated retraction. The joint diagonalization problem is used to demonstrate the performance of the algorithm with various parameters, line search conditions, and pairs of retraction and vector transport.

1 Introduction

In the Euclidean setting, the BFGS method is widely viewed as the best quasi-Newton method for solving smooth unconstrained optimization problems [5, 12]. Its global and superlinear local convergence is well understood for convex problems (see [5] and references therein). However, for nonconvex problems, its convergence properties are more intricate. Recently, Dai [4] has produced a nonconvex cost function for which the standard BFGS method does not converge. Modified BFGS methods exist that converge globally to critical points of nonconvex cost functions [10, 11].

Many Riemannian versions of the BFGS method have appeared, [6, 9, 13–15], but complete global and local convergence analyses that are not restricted to a specific cost function or a manifold are only given in two of them [9, 13]. The

W. Huang (✉) • P.-A. Absil
Department of Mathematical Engineering, ICTEAM Institute, Université catholique de Louvain, Louvain-la-Neuve, Belgium
e-mail: huwst08@gmail.com; absil@inma.ucl.ac.be

K.A. Gallivan
Department of Mathematics, Florida State University, Tallahassee, FL, USA
e-mail: kgallivan@fsu.edu

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_60

analyses of both methods require the cost function to satisfy a Riemannian version of convexity for global and superlinear local convergence.

In this paper, we generalize to manifolds the approach in [11] for nonconvex problems by using a Riemannian version of the cautious update of the Hessian approximation, and additionally a weak line search condition [3, (3.2), (3.3)]. Global and local superlinear convergence results are stated and the joint diagonalization problem [16] is used as an example to demonstrate numerical performance.

A key advantage of the proposed method over those in [9, 13] is that it offers more leeway on the choice of the vector transport. The version in [13] requires vector transport by differentiated retraction, which may not be available to users or may be too expensive. The version in [9] requires only the action of the differentiated retraction along a particular direction. In fact, any method that uses the Riemannian second Wolfe condition will require at least the action of the differentiated retraction along some particular direction. The proposed method is even less demanding: it no longer requires the second Wolfe condition, and the differentiated retraction can be completely avoided.

This paper is organized as follows. Section 2 presents notation used in this paper. Section 3 defines the Riemannian version of BFGS. Global and local convergence results are stated in Sect. 4. Numerical experiments are reported in Sect. 5.

2 Notation

The underlying concepts of Riemannian geometry can be found, e.g., in [1, 2]. We follow the notation of [1]. Let f denote a cost function defined on a d-dimensional Riemannian manifold $\mathscr{M}$ with the Riemannian metric $g : (\eta_x, \xi_x) \mapsto g_x(\eta_x, \xi_x) \in \mathbb{R}$. $\mathrm{T}_x \mathscr{M}$ denotes the tangent space of $\mathscr{M}$ at x and $\mathrm{T}\mathscr{M}$ denotes the tangent bundle, i.e., the set of all tangent spaces. For any $\eta_x \in \mathrm{T}_x \mathscr{M}$, $\eta_x^\flat$ denotes the function such that $\eta_x^\flat : \mathrm{T}_x \mathscr{M} \to \mathbb{R} : \xi_x \mapsto g_x(\eta_x, \xi_x)$.

A retraction is a C^1 map $R : \mathrm{T}\mathscr{M} \to \mathscr{M}$ such that $R(0_x) = x$ for all $x \in \mathscr{M}$ and $\frac{d}{dt}R(t\xi_x)|_{t=0} = \xi_x$ for all $\xi_x \in \mathrm{T}_x \mathscr{M}$. The domain of R does not have to be the entire tangent bundle, however, it is usually the case in practice. In this paper, we assume that R is well-defined whenever needed. R_x denotes the restriction of R to $\mathrm{T}_x \mathscr{M}$. A vector transport $\mathscr{T} : \mathrm{T}\mathscr{M} \oplus \mathrm{T}\mathscr{M} \to \mathrm{T}\mathscr{M}, (\eta_x, \xi_x) \mapsto \mathscr{T}_{\eta_x}\xi_x$ with associated retraction R is a mapping[1] such that, for all (x, η_x) in the domain of R and all $\xi_x, \zeta_x \in \mathrm{T}_x \mathscr{M}$, it holds that (i) $\mathscr{T}_{\eta_x}\xi_x \in \mathrm{T}_{R(\eta_x)} \mathscr{M}$, (ii) $\mathscr{T}_{\eta_x}$ is a linear map. An isometric vector transport $\mathscr{T}_{\mathrm{S}}$ additionally satisfies $g_{R_x(\eta_x)}(\mathscr{T}_{\mathrm{S}_{\eta_x}}\xi_x, \mathscr{T}_{\mathrm{S}_{\eta_x}}\zeta_x) = g_x(\xi_x, \zeta_x)$. The vector transport by differentiated retraction $\mathscr{T}_R$ is defined to be $\mathscr{T}_{R_{\eta_x}}\xi_x := \frac{d}{dt}R_x(\eta_x + t\xi_x)|_{t=0}$.

[1] This mapping is not even required to be continuous in the definition. The smoothness is imposed in the convergence analyses.

3 Riemannian BFGS Method with Cautious Update

The proposed Riemannian generalization of the BFGS method with cautious update is stated in Algorithm 1.

When $\mathscr{M}$ is a Euclidean space, the line search condition in Step 4 of Algorithm 1 is weak since it has been shown in [3, Sections 3 and 4] that many line search conditions, e.g., the Curry-Altman condition, the Goldstein condition, the Wolfe condition and the Armijo-Goldstein condition, imply either (1) or (2) if the gradient of the function is Lipschitz continuous. In the Riemannian setting, note that the function $f \circ R_x : \mathrm{T}_x \mathscr{M} \to \mathbb{R}$ is defined on a linear space. It follows that the Euclidean results about line search are applicable, i.e., the above conditions also imply either (1) or (2) when the gradient of the function satisfies the Riemannian Lipschitz continuous condition [1, Definition 7.4.1].

Among several possible Riemannian generalizations of the BFGS update formula [9, 13, 15], we opt here for $\mathscr{B}_{k+1} = \tilde{\mathscr{B}} - \frac{\tilde{\mathscr{B}}_k s_k (\tilde{\mathscr{B}}_k^* s_k)^\flat}{(\tilde{\mathscr{B}}_k^* s_k)^\flat s_k} + \frac{y_k y_k^\flat}{y_k^\flat s_k}$, where $\tilde{\mathscr{B}}_k = \mathscr{T}_{\mathrm{S}_{\alpha_k \eta_k}} \circ \mathscr{B}_k \circ \mathscr{T}_{\mathrm{S}_{\alpha_k \eta_k}}^{-1}$, $y_k = \beta_k^{-1} \operatorname{grad} f(x_{k+1}) - \mathscr{T}_{\mathrm{S}_{\alpha_k \eta_k}} \operatorname{grad} f(x_k)$, $s_k = \mathscr{T}_{\mathrm{S}_{\alpha_k \eta_k}} \alpha_k \eta_k$, and β_k is an arbitrary number satisfying $|\beta_k - 1| \leq L_\beta \|\alpha_k \eta_k\|$ and $L_\beta > 0$ is a constant. The motivation for introducing β_k is to make this update subsume the update in [9], which uses $\beta_k = \frac{\|\alpha_k \eta_k\|}{\|\mathscr{T}_{R_{\alpha_k \eta_k}} \alpha_k \eta_k\|}$.

If $y_k^\flat s_k > 0$, then the symmetric positive definiteness of $\tilde{\mathscr{B}}_k$ implies the symmetric positive definiteness of $\mathscr{B}_{k+1}$ [9]. The positive definiteness of the sequence $\{\mathscr{B}_k\}$

Algorithm 1 Cautious RBFGS method

```
Input: Riemannian manifold $\mathscr{M}$ with Riemannian metric $g$; a retraction $R$; isometric vector
    transport $\mathscr{T}_{\mathrm{S}}$, with $R$ as the associated retraction; continuously differentiable real-valued
    function $f$ on $\mathscr{M}$, bounded below; initial iterate $x_0 \in \mathscr{M}$; initial Hessian approximation $\mathscr{B}_0$
    that is symmetric positive definite with respect to the metric $g$; convergence tolerance $\varepsilon > 0$;
    constants $\chi_1 > 0$ and $\chi_2 > 0$ in the line search condition;
1: $k \leftarrow 0$;
2: while $\|\operatorname{grad} f(x_k)\| > \varepsilon$ do
3:     Obtain $\eta_k \in \mathrm{T}_{x_k} \mathscr{M}$ by solving $\mathscr{B}_k \eta_k = -\operatorname{grad} f(x_k)$;
4:     Set $x_{k+1} = R_{x_k}(\alpha_k \eta_k)$, where $\alpha_k > 0$ is computed from a line search procedure to satisfy
       either
           $h_k(\alpha_k) - h_k(0) \leq -\chi_1 \frac{h_k'(0)^2}{\|\eta_k\|^2}$        (1)
       or
           $h_k(\alpha_k) - h_k(0) \leq \chi_2 h_k'(0),$        (2)
       where $h_k(t) = f(R_{x_k}(t\eta_k))$.
5:     Define the linear operator $\mathscr{B}_{k+1} : \mathrm{T}_{x_{k+1}} \mathscr{M} \to \mathrm{T}_{x_{k+1}} \mathscr{M}$ by (3);
6:     $k \leftarrow k + 1$;
7: end while
```

is important in the sense that it guarantees that the search direction is a descent direction. However, not all line search conditions imply $y_k^\flat s_k > 0$. In the existing papers [9, 13], the Wolfe condition with information about $\mathscr{T}_R$ is used to guarantee $y_k^\flat s_k > 0$. In this paper, instead of enforcing $y_k^\flat s_k > 0$ by the Wolfe condition, we guarantee symmetric positive definiteness of $\mathscr{B}_{k+1}$ by resorting to the following cautious update formula

$$\mathscr{B}_{k+1} = \begin{cases} \tilde{\mathscr{B}}_k - \frac{\tilde{\mathscr{B}}_k s_k (\tilde{\mathscr{B}}_k^* s_k)^\flat}{(\tilde{\mathscr{B}}_k^* s_k)^\flat s_k} + \frac{y_k y_k^\flat}{y_k^\flat s_k}, & \text{if } \frac{y_k^\flat s_k}{\|s_k\|^2} \geq \vartheta(\| \operatorname{grad} f(x_k) \|) \\ \tilde{\mathscr{B}}_k, & \text{otherwise,} \end{cases} \tag{3}$$

where ϑ is a monotone increasing function satisfying $\vartheta(0) = 0$ and ϑ strictly increasing at 0. Formula (3) reduces to the cautious update formula of [11] when $\mathscr{M}$ is a Euclidean space. Using update (3) does not require the Wolfe condition, which yields more leeway for choosing a line search condition. When $\frac{y_k^\flat s_k}{\|s_k\|^2} \not\geq \vartheta(\| \operatorname{grad} f(x_k)\|)$, $\mathscr{B}_{k+1}$ can be set to be any given constant matrix, e.g., id, rather than $\tilde{\mathscr{B}}_k$. The choice does not affect the theoretical results given later.

4 Convergence Analysis

Due to length limitations, we only state the convergence results without proofs. The proofs will be given in a forthcoming paper. Theorems 1 and 2 state the global and local convergence results respectively.

Theorem 1 *Let $\{x_k\}$ be a sequence generated by Algorithm 1. Assume that the level set $\Omega = \{x \in \mathscr{M} \mid f(x) \leq f(x_0)\}$ is compact, that there exists $L_1 > 0$ such that $\|\mathscr{T}_\eta \operatorname{grad} f(x) - \operatorname{grad} f(R_x(\eta))\| \leq L_1 \|\eta\|$ for all and η such that, and that the function $\hat{f} = f \circ R$ is radially L-C^1 function* [1, Definition 7.4.1] *for all $x \in \Omega$. Then* $\liminf_{k\to\infty} \| \operatorname{grad} f(x_k)\| = 0$.

Theorem 2 *Let $\{x_k\}$ be a sequence generated by Algorithm 1 that converges to a nondegenerate minimizer x^* of f. Suppose there exists a neighborhood $\tilde{\Omega}$ of x^* such that*

1. *the objective function f is twice continuously differentiable in $\tilde{\Omega}$ and there exists positive constants a_{10} and a_{11} such that for all $y \in \tilde{\Omega}$, $\| \operatorname{Hess} f(y) - \mathscr{T}_{S_\eta} \operatorname{Hess} f(x^*) \mathscr{T}_{S_\eta}^{-1} \| \leq a_{10} \|\eta\|$, where $\eta = R_{x^*}^{-1} y$;*
2. *the retraction R is twice continuously differentiable in $\tilde{\Omega}$ and there is a constant a_5 such that for all $x, y \in \tilde{\Omega}$, $\max_{t\in[0,1]} \operatorname{dist}(R_x(t\eta), x^*) \leq a_9 \max(\operatorname{dist}(x, x^*), \operatorname{dist}(y, x^*))$, where $\eta = R_x^{-1} y$;*
3. *the isometric vector transport $\mathscr{T}_S$ with associated retraction R is continuous and satisfies $\mathscr{T}_{0_x} \xi_x = \xi_x$ for all $\xi_x \in \mathrm{T}_x \mathscr{M}$, $\|\mathscr{T}_{S_\eta} - \mathscr{T}_{R_\eta}\| \leq \tilde{L}\|\eta\|$ and $\|\mathscr{T}_{S_\eta}^{-1} - \mathscr{T}_{R_\eta}^{-1}\| \leq \tilde{L}\|\eta\|$ for some constant $\tilde{L}$.*

Then there exists an index k_0 such that $\alpha_k = 1$ satisfies either (1) *or* (2) *for $k \geq k_0$. Moreover, if $\alpha_k = 1$ is used for all $k \geq k_0$, then x_k converges to x^* superlinearly, i.e.,* $\lim_{k\to\infty} \frac{\mathrm{dist}(x_{k+1},x^*)}{\mathrm{dist}(x_k,x^*)} = 0$.

It is shown in [7, Theorem 5.2.4] that $\alpha_k = 1$ eventually satisfies the two frequently used line search conditions, i.e., the Wolfe condition $h_k(\alpha_k) \leq h_k(0) + c_1\alpha_k h_k'(0)$ and $h_k'(\alpha_k) \geq c_2 h_k'(0)$, where $0 < c_1 < 0.5 < c_2 < 1$ and the Armijo-Goldstein condition $h_k(\alpha_k) \leq h_k(0) + \sigma\alpha_k h_k'(0)$, where α_k is the largest value in the set $\{t^{(i)} | t^{(i)} \in [\varrho_1 t^{(i-1)}, \varrho_2 t^{(i-1)}], t^{(0)} = 1\}$, $0 < \varrho_1 < \varrho_2 < 1$ and $0 < \sigma < 0.5$. Therefore, if $\alpha_k = 1$ is attempted first using one of the line search conditions, then the superlinear convergence of Algorithm 1 is obtained. At present, no conditions on χ_1 and χ_2 in (1) and (2) that guarantee a similar result are known.

If $h'(t)$ must be evaluated at $t \neq 0$ in line search conditions, such as the Wolfe condition, then the action of vector transport by differentiated retraction is required only in a particular direction. More specifically, $h'(t) = g_{R_{x_k}(t\eta_k)}(\mathrm{grad} f(R_{x_k}(t\eta_k)), \mathscr{T}_{R_{t\eta_k}}\eta_k)$ requires the action of vector transport by differentiated retraction, $\mathscr{T}_{R_\eta}\xi$, with η and ξ on a same direction. This is discussed in [9] and one approach to resort to as little information on the differentiated retraction as possible is also proposed. If $h'(t)$ is not required at $t \neq 0$, such as in the Armijo-Goldstein condition, then the differentiated retraction can be completely avoided since $\mathscr{T}_{R_{0\eta_k}}\eta_k = \eta_k$.

5 Experiments

In this section, we investigate numerically the impact of choosing the Wolfe versus the Armijo-Goldstein condition in Step 4 of on Algorithms 1.

5.1 *Problem, Retraction, Vector Transport and Step Size*

The joint diagonalization (JD) problem on the Stiefel manifold [16] is used to illustrate the numerical performance:

$$\min_{X\in\mathrm{St}(p,n)} f(X) = \min_{X\in\mathrm{St}(p,n)} -\sum_{i=1}^{N} \| \mathrm{diag}(X^T C_i X)\|_2^2,$$

where $\mathrm{St}(p,n) = \{X \in \mathbb{R}^{n\times p} | X^T X = I_p\}$, matrices $C_1, \ldots, C_N$ are given symmetric matrices, $\mathrm{diag}(M)$ denotes the vector formed by the diagonal entries of matrix M, and $\| \cdot \|_2$ denotes the 2-norm.

The Stiefel manifold $\mathrm{St}(p,n)$ can be viewed as a submanifold of $\mathbb{R}^{n\times p}$. The chosen Riemannian metric g on $\mathrm{St}(p,n)$ is the metric endowed from its embedding space, i.e., $g(\eta_X, \xi_X) = \operatorname{tr}(\eta_X^T \xi_X)$. With this Riemannian metric g, the gradient is given in [16, Section 2.3]. As discussed in [8, Section 2.2], a tangent vector $\eta_X \in \mathrm{T}_X \mathscr{M}$ can be represented by a vector in the embedding space $\mathbb{R}^{n\times p}$ or a d-dimensional coefficient vector of a basis of $\mathrm{T}_X \mathscr{M}$, where $d = np - p(p+1)/2$ is the dimension of $\mathrm{St}(p, n)$. In our experiments, we use a d-dimensional representation of tangent vectors. By varying the basis and fixing the coefficients, one can define the vector transport by parallelization [8, Section 2.3.1 and 5]. The implementation of vector transport is then simply an identity [7, Section 9.5].

The retraction is chosen to be qf retraction [1, (4.7)]

$$R_X(\eta_X) = \mathrm{qf}(X + \eta_X), \tag{4}$$

where qf denotes the Q factor of the QR decomposition with nonnegative elements on the diagonal of R.

5.2 *Tests and Results*

The C_i matrices are selected as $C_i = R_i + R_i^T$, where the elements of $R_i \in \mathbb{R}^{n\times n}$ are independently drawn from the standard normal distribution. The initial iterate X_0 is given by applying Matlab's function *orth* to a matrix whose elements are drawn from the standard normal distribution using Matlab's *randn*. The code can be found in http://www.math.fsu.edu/~whuang2/papers/ARBMNOP.htm.

Let RBFGS-W and RBFGS-A denote Algorithm 1 with the Wolfe condition and the Armijo-Goldstein condition respectively. Since the Wolfe condition requires the evaluation of $h'(t)$ at $t \neq 0$, we use the locking condition proposed in [9], which restricts the retraction R and the isometric vector transport $\mathscr{T}_{\mathrm{S}}$:

$$\mathscr{T}_{\mathrm{S}_\xi}\xi = \beta \mathscr{T}_{R_\xi}\xi, \quad \beta = \frac{\|\xi\|}{\|\mathscr{T}_{R_\xi}\xi\|}. \tag{5}$$

Let RV1 denote retraction (4) and the vector transport by parallelization, which does not satisfy the locking condition (5); RV2 denote retraction (4) and the vector transport using the approach of [9, Section 4.2], which does satisfy the locking condition (5) but the vector transport is not smooth and relatively expensive.

The experimental results with various parameters and algorithms are reported in Table 1. Note that there is no result for RBFGS-W with RV1 since the well-definedness of RBFGS-W requires the locking condition. It can be seen that the performances of the Armijo-Goldstein condition and the Wolfe condition with the chosen algorithms are similar.

RBFGS with RV1 performs worse than RBFGS with RV2 in the sense of number of function and gradient evaluations. This implies that the locking condition, to some extent, reduces the number of function and gradient evaluations in RBFGS

Table 1 An average of 1000 random runs of RBFGS. $n = 12$, $p = 8$, $c_1 = \sigma = 10^{-4}$. *iter*, *nf*, *ng*, *nV* and *t* denote the number of iterations, number of function evaluations, number of gradient evaluations, number of vector transport and computational time (millisecond) respectively

	N		Armijo-Goldstien: $[\varrho_1, \varrho_2]$				Wolfe: c_2			
			$[\frac{1}{2}, \frac{1}{2}]$	$[\frac{1}{4}, \frac{3}{4}]$	$[\frac{1}{16}, \frac{15}{16}]$	$[\frac{1}{64}, \frac{63}{64}]$	$\frac{1}{2}$	$\frac{3}{4}$	$\frac{15}{16}$	$\frac{63}{64}$
RV1	128	*iter*	239	194	191	191	\	\	\	\
		nf	306	213	206	206	\	\	\	\
		ng	240	195	192	192	\	\	\	\
		nV	477	389	381	381	\	\	\	\
		t	31.8	26.3	25.8	26.1	\	\	\	\
	512	*iter*	196	191	191	191	\	\	\	\
		nf	215	208	208	207	\	\	\	\
		ng	197	192	192	192	\	\	\	\
		nV	392	382	383	383	\	\	\	\
		t	93.2	89.6	88.8	89.2	\	\	\	\
RV2	128	*iter*	146	164	167	147	123	132	136	142
		nf	170	197	203	168	186	184	168	165
		ng	147	165	168	148	167	162	150	147
		nV	293	327	335	294	413	422	420	426
		t	26.4	28.9	29.4	26.4	28.0	27.6	26.0	25.6
	512	*iter*	149	149	153	148	131	138	140	151
		nf	169	169	175	166	197	189	171	180
		ng	150	150	154	149	176	166	153	156
		nV	298	299	305	296	434	436	431	449
		t	78.2	77.6	78.9	77.5	89.2	84.7	79.1	80.0

with either the Armijo-Goldstein condition or the Wolfe condition. Note that even though $h'(t)$ at $t \neq 0$ is not used in the Armijo-Goldstein line search condition, the locking condition can still reduce the number of function and gradient evaluations. However, due to the low complexities on vector transport, RBFGS-A with RV1 still have competitive performance in the sense of computational time.

6 Conclusion

The results demonstrate the global convergence expected in the algorithm. While the locking condition is no longer required, we see that using it reduces the number of function and gradient evaluations. For problems such as joint diagonalization with large enough N so those evaluations are dominated computationally, a reduction in overall time results.

Acknowledgements This paper presents research results of the Belgian Network DYSCO (Dynamical Systems, Control, and Optimization), funded by the Interuniversity Attraction Poles Programme initiated by the Belgian Science Policy Office. This work was supported by grant FNRS PDR T.0173.13.

References

1. P.-A. Absil, R. Mahony, R. Sepulchre, *Optimization Algorithms on Matrix Manifolds* (Princeton University Press, Princeton, 2008)
2. W.M. Boothby, *An Introduction to Differentiable Manifolds and Riemannian Geometry*, 2nd edn. (Academic, Orlando, 1986)
3. R.H. Byrd, J. Nocedal, A tool for the analysis of quasi-newton methods with application to unconstrained minimization. SIAM J. Numer. Anal. **26**(3), 727–739 (1989)
4. Y.-H. Dai, A perfect example for the BFGS method. Math. Program. **138**(1–2), 501–530 (2013). doi:10.1007/s10107-012-0522-2
5. J.E. Dennis, R.B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations* (Springer, New Jersey, 1983)
6. D. Gabay, Minimizing a differentiable function over a differential manifold. J. Optim. Theory Appl. **37**(2), 177–219 (1982)
7. W. Huang, Optimization algorithms on Riemannian manifolds with applications, Ph.D. thesis, Department of Mathematics, Florida State University (2013)
8. W. Huang, P.-A. Absil, K. A. Gallivan, A Riemannian symmetric rank-one trust-region method. Math. Program. **150**(2), 179–216 (2015)
9. W. Huang, K.A. Gallivan, P.-A. Absil, A Broyden class of quasi-Newton methods for Riemannian optimization. SIAM J. Optim. **25**(3), 1660–1685 (2015)
10. D.-H. Li, M. Fukushima, A modified BFGS method and its global convergence in nonconvex minimization. J. Comput. Appl. Math. **129**, 15–35 (2001)
11. D.-H. Li, M. Fukushima, On the global convergence of the BFGS method for nonconvex unconstrained optimization problems. SIAM J. Optim. **11**(4), 1054–1064 (2001). doi:10.1137/S1052623499354242
12. J. Nocedal, S.J. Wright, *Numerical Optimization*, 2nd edn. (Springer, New York, 2006)
13. W. Ring, B. Wirth, Optimization methods on Riemannian manifolds and their application to shape space. SIAM J. Optim. **22**(2), 596–627 (2012). doi:10.1137/11082885X
14. B. Savas, L.H. Lim, Quasi-Newton methods on Grassmannians and multilinear approximations of tensors. SIAM J. Sci. Comput. **32**(6), 3352–3393 (2010)
15. M. Seibert, M. Kleinsteuber, K. Hüper, Properties of the BFGS method on Riemannian manifolds. Mathematical System Theory – Festschrift in Honor of Uwe Helmke on the Occasion of his Sixtieth Birthday (2013), pp. 395–412
16. F.J. Theis, T.P. Cason, P.-A. Absil, Soft dimension reduction for ICA by joint diagonalization on the Stiefel manifold, in *Proceedings of the 8th International Conference on Independent Component Analysis and Signal Separation*, Paraty, vol. 5441 (2009), pp. 354–361

Discrete Lie Derivative

Marc Gerritsma, Jeroen Kunnen, and Boudewijn de Heij

Abstract Convection is an important transport mechanism in physics. Especially, in fluid dynamics at high Reynolds numbers this term dominates. Modern mimetic discretization methods consider physical variables as differential k-forms and their discrete analogues as k-cochains. Convection, in this parlance, is represented by the Lie derivative, $\mathscr{L}_X$. In this paper we design reduction operators, $\mathscr{R}$ from differential forms to cochains and define a discrete Lie derivative, L_X which acts on cochains such that the commutation relation $\mathscr{R}\mathscr{L}_X = \mathsf{L}_X\mathscr{R}$ holds.

1 Introduction

Differential forms have a natural discrete analogue in terms of cochains. The exterior derivative, d, which represents the gradient, curl and divergence operator is naturally represented at the discrete level by the coboundary operator, δ. If $\mathscr{R}$ denotes the De Rham or the reduction map, which converts continuous variables to discrete degrees of freedom, we have the commutation relation $\mathscr{R}\mathrm{d} = \delta\mathscr{R}$, see [1–3, 6, 9, 10, 13, 14].

This fully discrete representation of the exterior derivative is possible, because the exterior derivative is an *intrinsic operator*, which means that this operation is coordinate- and metric-free. The coboundary operator, likewise, does not depend on the mesh size, the shape of the mesh or the order of the numerical method, and is therefore referred to as a *topological operator*.

The Lie derivative, $\mathscr{L}_X$, which represents the rate of change of a differential form in the direction of a vector field X is also an intrinsic operator and therefore we expect that there exists a purely discrete operator L_X which operates on cochains and satisfies the commutation relation $\mathscr{R}\mathscr{L}_X = \mathsf{L}_X\mathscr{R}$.

Convective transport and the use of the Lie derivative of differential forms have been addressed in various papers, see, for instance, [3, 4, 8, 12]. Most of these

M. Gerritsma (✉) • J. Kunnen • B. de Heij
Delft University of Technology, Kluyverweg 1, 2629 HS Delft, The Netherlands
e-mail: m.i.gerritsma@student.tudelft.nl; j.kunnen@student.tudelft.nl; b.deheij@student.tudelft.nl

B. Karasözen et al. (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*, Lecture Notes in Computational Science and Engineering 112, DOI 10.1007/978-3-319-39929-4_61

methods are based on Bossavit's extrusion idea, which requires the complete flow associated with the convective velocity field. In [13] the Lie derivative is evaluated by means of inner-products and musical operators[1] which are metric-dependent. In the framework of compatible discrete operator schemes, Cantin and Ern present a discrete contraction operator, [5].

The difficulty of constructing a fully discrete Lie derivative can be traced back to the dual interpretation of differential forms. A 1-form, for instance, can be interpreted as the "thing which occurs under integral signs", [7]. This interpretation leads to the discrete representation of 1-forms as 1-cochain, where the reduction operation consists of integrating the 1-form over a 1-chain. In the second interpretation a 1-form at a point p is an element of the cotangent space, $T_p^*(\mathscr{M})$, which is the space of linear functionals acting on the linear vector space $T_p(\mathscr{M})$. It is this second interpretation which allows the pointwise contraction with a vector field and the definition of a fully discrete Lie derivative.

2 Properties of the Lie Derivative

Let $X \in \mathfrak{X}(\mathscr{M})$ be a smooth vector field, then the Lie derivative of a k-form $\omega^{(k)}$ along the vector field will be denoted by $\mathscr{L}_X\omega^{(k)}$. The Lie derivative has the following properties:

The Lie derivative is linear w.r.t. differential forms, i.e. if $\alpha^{(k)}, \beta^{(k)} \in \Lambda^k(\mathscr{M})$ and $\lambda, \mu \in \mathbb{R}$ then

$$\mathscr{L}_X(\lambda\alpha^{(k)} + \mu\beta^{(k)}) = \lambda\mathscr{L}_X(\alpha^{(k)}) + \mu\mathscr{L}_X(\beta^{(k)}) \ .$$

The Lie derivative is linear in the vector fields, i.e. if $X, Y \in \mathfrak{X}(\mathscr{M})$ and $\lambda, \mu \in \mathbb{R}$ then

$$\mathscr{L}_{(\lambda X+\mu Y)}\omega^{(k)} = \lambda\mathscr{L}_X\omega^{(k)} + \mu\mathscr{L}_Y\omega^{(k)} \ .$$

The Lie derivative satisfies a Leibniz rule for the wedge product, i.e. if $\alpha^{(k)} \in \Lambda^k(\mathscr{M})$ and $\beta^{(l)} \in \Lambda^l(\mathscr{M})$, then $\alpha^{(k)} \wedge \beta^{(l)}$ is an $(k+l)$-form and

$$\mathscr{L}_X\left(\alpha^{(k)} \wedge \beta^{(l)}\right) = \left(\mathscr{L}_X\alpha^{(k)}\right) \wedge \beta^{(l)} + \alpha^{(k)} \wedge \left(\mathscr{L}_X\beta^{(l)}\right) \ .$$

[1] The vector space $T_p(\mathscr{M})$ and $T_p^*(\mathscr{M})$ are ismorphic, but there is no natural isomorphism. One way to associate vectors $\mathbf{v}$ to covectors α is by means of the metric tensor: $\alpha_i = g_{ij}v^j$. With this association we have $\mathbf{v}^\flat = \alpha$ and $\alpha^\sharp = \mathbf{v}$. By construction the musical operators are metric-dependent.

For a 0-form $f^{(0)}$ the Lie derivative can be written in any of the equivalent forms

$$\mathscr{L}_X f^{(0)} = \iota_X \mathrm{d} f^{(0)} = \langle \mathrm{d} f^{(0)}, X\rangle = X[f] \;, \tag{1}$$

where ι_X is the interior product, d the exterior derivative and $\langle\cdot,\cdot\rangle$ denotes point-wise duality pairing between a covector, $\mathrm{d}f^{(0)}$, and a vector, X. For general k-forms $\omega^{(k)}$ the Lie derivative can be evaluated by Cartan's magic formula

$$\mathscr{L}_X \omega^{(k)} = \mathrm{d}\iota_X \omega^{(k)} + \iota_X \mathrm{d}\omega^{(k)} \;.$$

3 Derivations at a Point

Consider a point $p \in \mathscr{M}$ and a convex neighborhood U_p of p. Let C_p^m be the set of all C^m functions on U_p. A linear map $D : C_p^m \to \mathbb{R}$ is called a *point derivation* if for all $f, h \in C_p^m$ it satisfies the Leibniz rule

$$D(f \cdot h)_p = D(f)_p \cdot h(p) + f(p) \cdot D(h)_p \in \mathbb{R} \;.$$

The set of all point derivations D_p at a point $p \in \mathscr{M}$ form a linear vector space, $\mathscr{D}_p(\mathscr{M})$, and it can be shown that this linear vector is isomorphic to the tangent space $T_p(\mathscr{M})$, see [15].

When $f \in C_p^m$, $m \geq 1$, there exist functions $g_1(x), \ldots, g_d(x) \in C_p^{m-1}$ such that

$$f(x) = f(p) + \sum_{i=1}^{d} (x^i - p^i) g_i(x) \;, \quad x \in U_p \;, \tag{2}$$

see [15, Lemma 1.4] or [11, §3.4]. If $\mathbf{v}_p = c^1(\partial_1)_p + \ldots + c^d(\partial_d)_p$ is a vector defined at $p \in \mathscr{M}$, then

$$\begin{aligned}\mathbf{v}_p[f] &= \mathbf{v}_p[f(p)] + \sum_{i=1}^{d} \mathbf{v}_p[(x^i - p^i) g_i(x)] \\ &= 0 + \sum_{i=1}^{d} \mathbf{v}_p[x^i - p^i] g_i(p) + (x^i - p^i)_p \mathbf{v}_p[g] = \sum_{i=1}^{d} c^i g_i(p) \;,\end{aligned} \tag{3}$$

where we used $\mathbf{v}[x^i] = c^i$. So a point-wise exact evaluation of the Lie derivative is possible if the vector field is defined in the points p and the function values of $g_i(x)$ are known at the points p. In conventional mimetic methods such as [1, 6, 10], the reduction or De Rham map of a 0-form (functions) is done by only evaluating the value of the function in points, i.e. $\mathscr{R}_p(f) = f(p)$. However, (3) reveals that the value of f at the point p is irrelevant. For a discrete Lie derivative more information is necessary and therefore the newly proposed reduction operator evaluates f *and* the functions $g_i(x)$ in the points. If $\mathscr{R}_p^{(1)}$ denotes the reduction of $f^{(0)}$ in a point p, it is given by

$$\mathscr{R}_p^{(1)}(f) := (f_p, g_{1,p}, \ldots, g_{d,p}) \ , \tag{4}$$

where $f_p = f(p)$ and $g_{i,p} = g_i(p)$. So instead of associating one value to a point, we now associate with all points in the mesh a vector of length $(1 + d)$. We will refer to such a vector in a point as a *vector-valued* 0*-cochain.*

The conventional reduction, in which only the functional value at the point p is sampled will be referred to as $\mathscr{R}_p^{(0)}(f) = f_p = f(p)$. If the vector field, $\mathbf{v}$ is known at all points p, then the following diagram commutes

$$\begin{tikzcd}
f \arrow[r, "\mathscr{L}_{\mathbf{v}}"] \arrow[d, "\mathscr{R}_p^{(1)}"'] & \mathscr{L}_{\mathbf{v}} f \arrow[d, "\mathscr{R}_p^{(0)}"] \\
(f_p, g_{1,p}, \ldots, g_{n,p}) \arrow[r, "\mathsf{L}_{\mathbf{v}}"] & c^1 g_{1,p} + \ldots + c^n g_{n,p}
\end{tikzcd}$$

That is, we have for all $f \in C^m(\Omega)$, $m \geq 1$

$$\mathscr{R}_p^{(0)}(\mathscr{L}_{\mathbf{v}} f) = \mathsf{L}_{\mathbf{v}} \mathscr{R}_p^{(1)}(f) \ . \tag{5}$$

The commutating diagram which defines the discrete Lie derivative applied to 0-forms, is graphically represented in Fig. 1, where $f(x) = \sin(2\pi x) + 3\cos(3\pi x)$ and $\mathbf{v}(x) = 0.5\sin(14\pi x)\partial_x$. Note that the reduction for f is $\mathscr{R}_p^{(1)}$, i.e. it samples function values and derivatives, while the reduction of the $\mathscr{L}_{\mathbf{v}} f$ is done with the 'conventional' $\mathscr{R}_p^{(0)}$ which only evaluates the value in a point. If we take the exterior derivative of (3) we have

$$\mathrm{d}f(x) = \sum_{i=1}^{N} \left[(x^i - p^i)\mathrm{d}g_i(x) + g_i(x)\mathrm{d}x^i \right] \ .$$

This allows us to define the exterior derivative in a point p, by setting $x = p$

$$\mathrm{d}f|_p = g_i(p)\mathrm{d}x^i \ . \tag{6}$$

This implies that $c^1 g_{1,p} + \ldots + c^n g_{n,p}$ is the point-wise contraction of $\mathbf{v} = v^i \partial_{x^i}$ and the point-wise exterior derivative (6), which is one of the forms of the Lie derivative given in (1). It follows that if $\mathbf{v}$ and $\mathbf{w}$ are vector fields that this construction gives the pointwise $\mathsf{L}_{(\mathbf{v}+\mathbf{w})} f = \mathsf{L}_{\mathbf{v}} f + \mathsf{L}_{\mathbf{w}} f$ and $\mathsf{L}_{\mathbf{v}}(f + h) = \mathsf{L}_{\mathbf{v}} f + \mathsf{L}_{\mathbf{v}} h$. It remains to show that this construction satisfies the Leibniz rule.

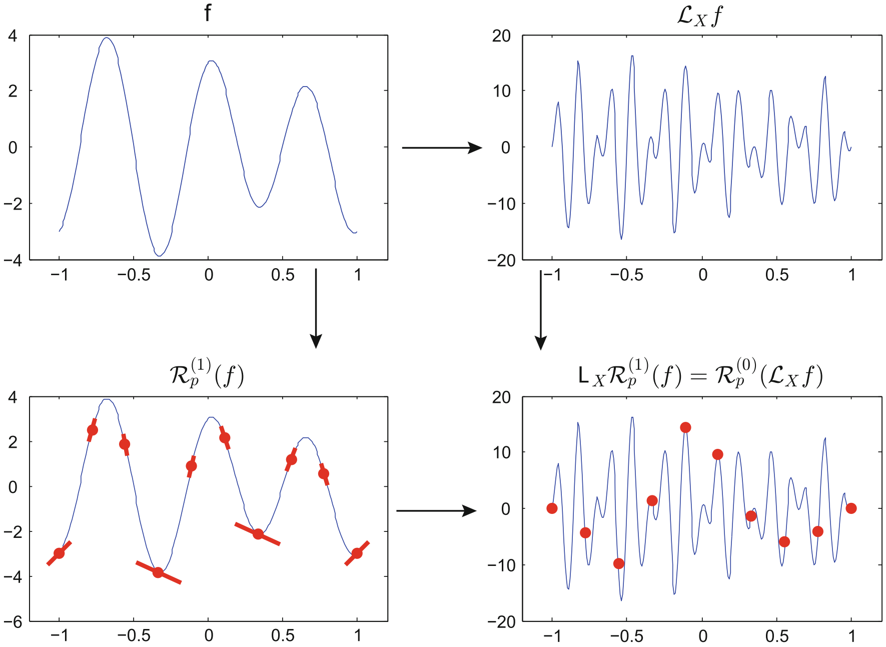


Fig. 1 Graphical representation of the commutation relation (5). In the *top row* the continuous Lie derivative is applied to the f. In the *bottom row* the discrete Lie derivative is applied to the vector-valued 0-cochain (4). The reduction $\mathscr{R}_p^{(1)}$ samples the function value and the local gradient, while $\mathscr{R}_p^{(0)}$ only samples the function values

Assume that $f(x)$ and $h(x)$ can be written in the form (2) for $x \in U_p$, i.e. for $f, h \in C_p^m$ there exist functions $g_i, k_i \in C_p^{m-1}$ such that

$$f(x) = f(p) + \sum_{i=1}^{d}(x^i - p^i)g_i(x) \ , \quad h(x) = h(p) + \sum_{i=1}^{d}(x^i - p^i)k_i(x) \ ,$$

for $x \in U_p$. Then we have

$$\begin{aligned} f(x) \cdot h(x) &= f(p) \cdot h(p) \\ &\quad + \sum_{i=1}^{d}(x^i - p^i)\left[f(p)k_i(x) + g_i(x)h(p) + \sum_{j=1}^{d}(x^j - p^j)g_i(x)k_j(x) \right] \end{aligned}$$

Then the reduction $\mathscr{R}_p^{(1)}(f \cdot h)$ is given by

$$\mathscr{R}_p^{(1)}(f \cdot h) = (f(p)h(p), f(p)k_1(p) + g_1(p)h(p), \ldots, f(p)k_d(p) + g_d(p)h(p)) \ . \qquad (7)$$

The reduction (7) defines the product of vector-valued cochains which turns the linear space of vector-valued 0-cochains into an algebra. When we apply the discrete Lie derivative to (7), we obtain

$$\mathsf{L}_{\mathbf{v}}\mathscr{R}_p^{(1)}(fh) = f(p)\cdot\mathsf{L}_{\mathbf{v}}\mathscr{R}_p^{(1)}(h) + \mathsf{L}_{\mathbf{v}}\mathscr{R}_p^{(1)}(f)\cdot h(p)\ ,$$

therefore the discrete Lie derivative satisfies the Leibniz rule.

4 The Discrete Lie Derivative of a 1-Cochain

Let $a^{(1)} = a_i(x)\mathrm{d}x^i$ be a differential 1-form, then the conventional reduction, denoted by $\mathscr{R}_e^{(0)}$ associates with each edge e in the cell complex the value

$$\mathscr{R}_e^{(0)}(a^{(1)}) := \int_e a^{(1)}\ ,$$

where the subscript e in the reduction operator indicates that it reduces the 1-form to values associated with edges. For the extended reduction, denoted by $\mathscr{R}_{pe}^{(1)}$, we will reduce the 1-form at the nodes *and* at the edges

$$\mathscr{R}_{pe}^{(1)}(a^{(1)}) = (a_{1,p},\ldots,a_{d,p},\bar{a}_e)\ , \tag{8}$$

where $\bar{a}_e = \mathscr{R}_e^{(0)}(a^{(1)})$ and $a_{i,p} = a_i(p)$. So in this reduction we have the point-wise evaluation of the functions $a_i(x)$ and the integral value along the edges of the grid. For an exact 1-form $a^{(1)} = \mathrm{d}f$, the functions $a_i(x)$ are given by $g_i(x)$ according to (6). Contraction of this reduced 1-form with a vector defined at p is then given by

$$\iota_{\mathbf{v}}|_p\,\mathscr{R}_{pe}^{(1)}(a^{(1)}) = c^1a_1(p)+\ldots+c^da_d(p) = \langle\mathbf{v},a_i(x)\mathrm{d}x^i\rangle_p\ . \tag{9}$$

The discrete Lie derivative applied to reduced 1-forms then follows from the application of the coboundary operator, δ, to (9). So to every edge we assign the value $\langle\mathbf{v},a_i(x)\mathrm{d}x^i\rangle_{p+1} - \langle\mathbf{v},a_i(x)\mathrm{d}x^i\rangle_p$, where $(p+1)$ and p denote the boundary points of the edge under consideration. This construction implies that the following diagram commutes.

$$\begin{tikzcd}
a^{(1)} \arrow[r, "\iota_{\mathbf{v}}"] \arrow[d, "\mathscr{R}_{pe}^{(1)}"'] & \iota_{\mathbf{v}}a^{(1)} \arrow[r, "\mathrm{d}"] \arrow[d, "\mathscr{R}_p^{(0)}"] & \mathscr{L}_{\mathbf{v}}a^{(1)} \arrow[d, "\mathscr{R}_e^{(0)}"] \\
(a_{1,p},\ldots,a_{d,p},\bar{a}_e) \arrow[r, "\iota_{\mathbf{v}}|_p"] & c^1a_1(p)+\ldots+c^da_d(p) \arrow[r, "\delta"] & \mathsf{L}_{\mathbf{v}}\mathscr{R}_{pe}^{(1)}(a^{(1)})
\end{tikzcd} \tag{10}$$

This commutation relation is graphically depicted in Fig. 2, where we used $a^{(1)} = \mathrm{d}f$ with the same vector field and the same f as in the previous test case. In the lower left

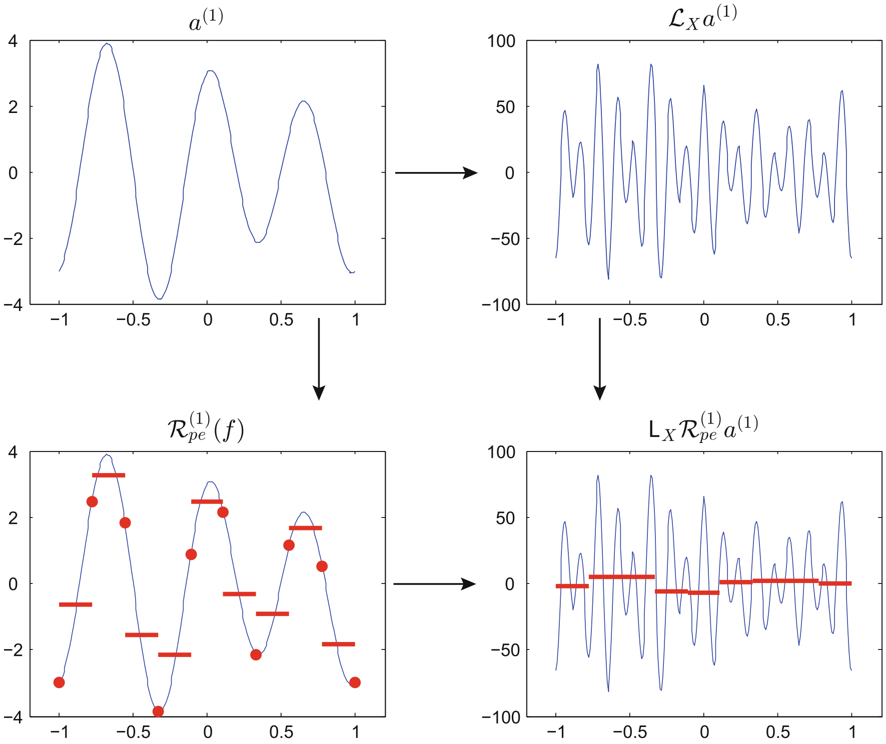


Fig. 2 Graphical representation of the commutation relation (10). In the *top row* the continuous Lie derivative is applied to $a^{(1)}$. In the *bottom row* the discrete Lie derivative is applied to the reduced 1-form $\mathscr{R}_{pe}^{(1)}(a^{(1)})$. This reduction samples function values (*red dots*) and integral values (*red bars*)

plot in this figure, we see that we sample $a(x)\mathrm{d}x$ in the points (red dots) and evaluate the integral over the edges (red bars). The integrals in the lower right plot are exact. While the reduction in Fig. 2 was performed in 10 points and along 9 edges, Fig. 3 shows the same discrete operation in 100 points and along 99 edges. Note that in the one-dimensional case, the value assigned to an edge equals the flux on the right, $\langle \mathbf{v}, a_i(x)\mathrm{d}x^i\rangle_{p+1}$ minus the flux on the left $\langle \mathbf{v}, a_i(x)\mathrm{d}x^i\rangle_p$. This construction resembles the one used in finite volume methods. The current method differs from conventional finite volume methods in that the value of a is available at the cell interfaces, whereas in finite volume methods this value needs to be reconstructed from the cell averages. Such a reconstruction inevitably requires approximation and leads to error in the approximation of the Lie derivative. The current approach avoids the approximation step and is therefore exact.

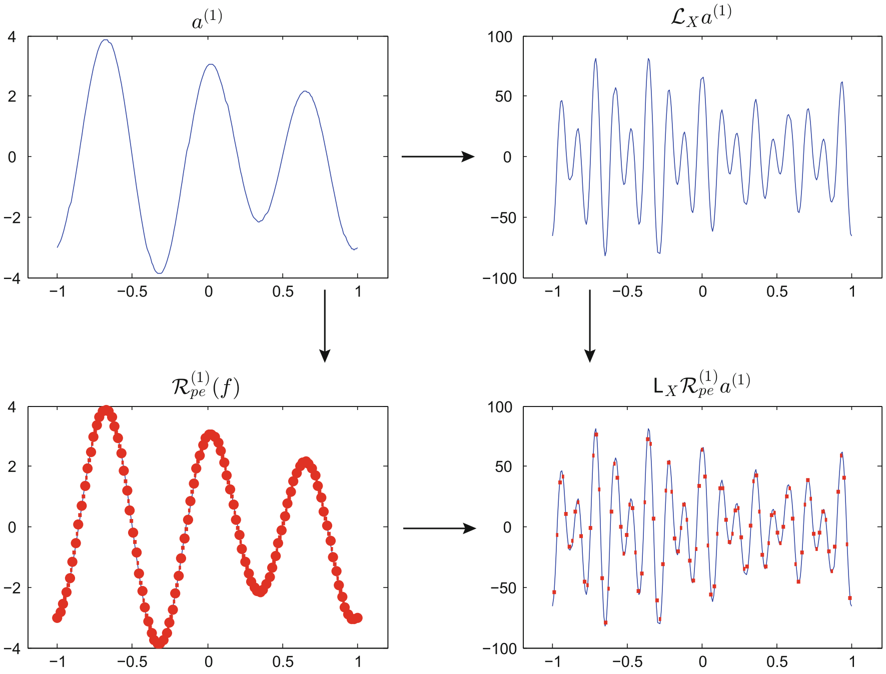


Fig. 3 Graphical representation of the commutation relation (10), where we used 100 points and 99 edges

5 Conclusion

This paper introduces a fully discrete Lie derivative and outlines the construction for 0-forms and 1-forms. This requires new reduction operators for both 0-forms and 1-forms. To complete the discretization method reconstruction operators, $\mathscr{I}$, need to be defined which commute with the discrete Lie derivative, [1, 10]. These operators will be addressed in future work.

References

1. P. Bochev, J. Hyman, Principles of mimetic discretizations of differential operators, in *Compatible Spatial Discretizations*, ed. by D. Arnold, P. Bochev, R. Nicolaides, M. Shashkov. The IMA Volumes in Mathematics and Its Applications, vol. 42 (Springer, New York, 2006), pp. 89–119
2. J. Bonelle, A. Ern, Analysis of compatible discrete operator schemes for elliptic problems on polyhedral meshes. ESIAM: Math. Model. Numer. Anal. **48**(2), 553–581 (2014)
3. A. Bossavit, Extrusion, contraction: their discretization via Whitney forms. COMPEL **22**(3), 470–480 (2002)

4. A. Bossavit, in *Applied Differential Geometry* (2005). http://butler.cc.tut.fi/~bossavit/BackupICM/Compendium.html
5. P. Cantin, A. Ern, Vertex-based compatible discrete operator schemes on polyhedral meshes for advection-diffusion equations, HAL Id: hal-01141290. https://hal.archives-ouvertes.fr/hal-00141290v2, to appear in Comput. Methods Appl. Math. 2016
6. M. Desbrun, A.N. Hirani, M. Leok, J.E. Marsden, Discrete exterior calculus (2005). Arxiv preprint math/0508341
7. Flanders, *Differential Forms with Applications to the Physical Sciences* (Dover books, New York, 1963)
8. H. Heumann, R. Hiptmair, K. Li, J. Xu, Fully discrete semi-Lagrangian methods for advection of differential forms. BIT Numer. Math. **52**, 981–1007 (2012)
9. J. Kreeft, M. Gerritsma, Mixed mimetic spectral element methods for Stokes flow: a pointwise divergence-free solution. J. Comput. Phys. **240**, 284–309 (2013)
10. J. Kreeft, A. Palha, M. Gerritsma, Mimetic framework on curvilinear quadrilaterals of arbitrary order (2011). arXiv:1111.4304
11. A. McInerney, *First Steps in Differential Geometry – Riemann, Contact, Symplectic* (Springer, New York/Dordrecht/Heidelberg/London, 2013)
12. P. Mullen, A. McKenzie, D. Pavlov, L. Durant, Y. Tong, E. Kanso, J.E. Marsden, M. Desbrun, Discrete Lie advection of differential forms. Found. Comput. Math. **11**(2), 131–149 (2011)
13. A. Palha, P.P. Rebelo, M. Gerritsma, Mimetic spectral element advection. Lect. Notes Comput. Sci. Eng. **95**, 325–335 (2014)
14. N. Robidoux, S. Steinberg, A discrete vector calculus in tensor grids. Comput. Methods Appl. Math. **1**, 1–44 (2011)
15. L.W. Tu, *An Introduction to Manifolds* (Springer, New York/Dordrecht/Heidelberg/London, 2011)

Editorial Policy

1. Volumes in the following three categories will be published in LNCSE:

i) Research monographs
ii) Tutorials
iii) Conference proceedings

Those considering a book which might be suitable for the series are strongly advised to contact the publisher or the series editors at an early stage.

2. Categories i) and ii). Tutorials are lecture notes typically arising via summer schools or similar events, which are used to teach graduate students. These categories will be emphasized by Lecture Notes in Computational Science and Engineering. **Submissions by interdisciplinary teams of authors are encouraged.** The goal is to report new developments – quickly, informally, and in a way that will make them accessible to non-specialists. In the evaluation of submissions timeliness of the work is an important criterion. Texts should be well-rounded, well-written and reasonably self-contained. In most cases the work will contain results of others as well as those of the author(s). In each case the author(s) should provide sufficient motivation, examples, and applications. In this respect, Ph.D. theses will usually be deemed unsuitable for the Lecture Notes series. Proposals for volumes in these categories should be submitted either to one of the series editors or to Springer-Verlag, Heidelberg, and will be refereed. A provisional judgement on the acceptability of a project can be based on partial information about the work: a detailed outline describing the contents of each chapter, the estimated length, a bibliography, and one or two sample chapters – or a first draft. A final decision whether to accept will rest on an evaluation of the completed work which should include

- at least 100 pages of text;
- a table of contents;
- an informative introduction perhaps with some historical remarks which should be accessible to readers unfamiliar with the topic treated;
- a subject index.

3. Category iii). Conference proceedings will be considered for publication provided that they are both of exceptional interest and devoted to a single topic. One (or more) expert participants will act as the scientific editor(s) of the volume. They select the papers which are suitable for inclusion and have them individually refereed as for a journal. Papers not closely related to the central topic are to be excluded. Organizers should contact the Editor for CSE at Springer at the planning stage, see *Addresses* below.

In exceptional cases some other multi-author-volumes may be considered in this category.

4. Only works in English will be considered. For evaluation purposes, manuscripts may be submitted in print or electronic form, in the latter case, preferably as pdf- or zipped ps-files. Authors are requested to use the LaTeX style files available from Springer at http://www.springer.com/gp/authors-editors/book-authors-editors/manuscript-preparation/5636 (Click on LaTeX Template → monographs or contributed books).

For categories ii) and iii) we strongly recommend that all contributions in a volume be written in the same LaTeX version, preferably LaTeX2e. Electronic material can be included if appropriate. Please contact the publisher.

Careful preparation of the manuscripts will help keep production time short besides ensuring satisfactory appearance of the finished book in print and online.

5. The following terms and conditions hold. Categories i), ii) and iii):

Authors receive 50 free copies of their book. No royalty is paid.
Volume editors receive a total of 50 free copies of their volume to be shared with authors, but no royalties.

Authors and volume editors are entitled to a discount of 33.3 % on the price of Springer books purchased for their personal use, if ordering directly from Springer.

6. Springer secures the copyright for each volume.

Addresses:

Timothy J. Barth
NASA Ames Research Center
NAS Division
Moffett Field, CA 94035, USA
barth@nas.nasa.gov

Michael Griebel
Institut für Numerische Simulation
der Universität Bonn
Wegelerstr. 6
53115 Bonn, Germany
griebel@ins.uni-bonn.de

David E. Keyes
Mathematical and Computer Sciences
and Engineering
King Abdullah University of Science
and Technology
P.O. Box 55455
Jeddah 21534, Saudi Arabia
david.keyes@kaust.edu.sa

and

Department of Applied Physics
and Applied Mathematics
Columbia University
500 W. 120 th Street
New York, NY 10027, USA
kd2112@columbia.edu

Risto M. Nieminen
Department of Applied Physics
Aalto University School of Science
and Technology
00076 Aalto, Finland
risto.nieminen@aalto.fi

Dirk Roose
Department of Computer Science
Katholieke Universiteit Leuven
Celestijnenlaan 200A
3001 Leuven-Heverlee, Belgium
dirk.roose@cs.kuleuven.be

Tamar Schlick
Department of Chemistry
and Courant Institute
of Mathematical Sciences
New York University
251 Mercer Street
New York, NY 10012, USA
schlick@nyu.edu

Editor for Computational Science
and Engineering at Springer:
Martin Peters
Springer-Verlag
Mathematics Editorial IV
Tiergartenstrasse 17
69121 Heidelberg, Germany
martin.peters@springer.com

Lecture Notes in Computational Science and Engineering

1. D. Funaro, *Spectral Elements for Transport-Dominated Equations.*

2. H.P. Langtangen, *Computational Partial Differential Equations.* Numerical Methods and Diffpack Programming.

3. W. Hackbusch, G. Wittum (eds.), *Multigrid Methods V.*

4. P. Deuflhard, J. Hermans, B. Leimkuhler, A.E. Mark, S. Reich, R.D. Skeel (eds.), *Computational Molecular Dynamics: Challenges, Methods, Ideas.*

5. D. Kröner, M. Ohlberger, C. Rohde (eds.), *An Introduction to Recent Developments in Theory and Numerics for Conservation Laws.*

6. S. Turek, *Efficient Solvers for Incompressible Flow Problems.* An Algorithmic and Computational Approach.

7. R. von Schwerin, _**M**ulti **B**ody **S**ystem **SIM**ulation._ Numerical Methods, Algorithms, and Software.

8. H.-J. Bungartz, F. Durst, C. Zenger (eds.), *High Performance Scientific and Engineering Computing.*

9. T.J. Barth, H. Deconinck (eds.), *High-Order Methods for Computational Physics.*

10. H.P. Langtangen, A.M. Bruaset, E. Quak (eds.), *Advances in Software Tools for Scientific Computing*.

11. B. Cockburn, G.E. Karniadakis, C.-W. Shu (eds.), *Discontinuous Galerkin Methods.* Theory, Computation and Applications.

12. U. van Rienen, *Numerical Methods in Computational Electrodynamics.* Linear Systems in Practical Applications.

13. B. Engquist, L. Johnsson, M. Hammill, F. Short (eds.), *Simulation and Visualization on the Grid.*

14. E. Dick, K. Riemslagh, J. Vierendeels (eds.), *Multigrid Methods VI.*

15. A. Frommer, T. Lippert, B. Medeke, K. Schilling (eds.), *Numerical Challenges in Lattice Quantum Chromodynamics.*

16. J. Lang, *Adaptive Multilevel Solution of Nonlinear Parabolic PDE Systems.* Theory, Algorithm, and Applications.

17. B.I. Wohlmuth, *Discretization Methods and Iterative Solvers Based on Domain Decomposition.*

18. U. van Rienen, M. Günther, D. Hecht (eds.), *Scientific Computing in Electrical Engineering.*

19. I. Babuška, P.G. Ciarlet, T. Miyoshi (eds.), *Mathematical Modeling and Numerical Simulation in Continuum Mechanics.*

20. T.J. Barth, T. Chan, R. Haimes (eds.), *Multiscale and Multiresolution Methods.* Theory and Applications.

21. M. Breuer, F. Durst, C. Zenger (eds.), *High Performance Scientific and Engineering Computing.*

22. K. Urban, *Wavelets in Numerical Simulation.* Problem Adapted Construction and Applications.

23. L.F. Pavarino, A. Toselli (eds.), *Recent Developments in Domain Decomposition Methods.*

24. T. Schlick, H.H. Gan (eds.), *Computational Methods for Macromolecules: Challenges and Applications.*

25. T.J. Barth, H. Deconinck (eds.), *Error Estimation and Adaptive Discretization Methods in Computational Fluid Dynamics*.

26. M. Griebel, M.A. Schweitzer (eds.), *Meshfree Methods for Partial Differential Equations*.

27. S. Müller, *Adaptive Multiscale Schemes for Conservation Laws*.

28. C. Carstensen, S. Funken, W. Hackbusch, R.H.W. Hoppe, P. Monk (eds.), *Computational Electromagnetics*.

29. M.A. Schweitzer, *A Parallel Multilevel Partition of Unity Method for Elliptic Partial Differential Equations*.

30. T. Biegler, O. Ghattas, M. Heinkenschloss, B. van Bloemen Waanders (eds.), *Large-Scale PDE-Constrained Optimization*.

31. M. Ainsworth, P. Davies, D. Duncan, P. Martin, B. Rynne (eds.), *Topics in Computational Wave Propagation*. Direct and Inverse Problems.

32. H. Emmerich, B. Nestler, M. Schreckenberg (eds.), *Interface and Transport Dynamics.* Computational Modelling.

33. H.P. Langtangen, A. Tveito (eds.), *Advanced Topics in Computational Partial Differential Equations.* Numerical Methods and Diffpack Programming.

34. V. John, *Large Eddy Simulation of Turbulent Incompressible Flows.* Analytical and Numerical Results for a Class of LES Models.

35. E. Bänsch (ed.), *Challenges in Scientific Computing - CISC 2002.*

36. B.N. Khoromskij, G. Wittum, *Numerical Solution of Elliptic Differential Equations by Reduction to the Interface.*

37. A. Iske, *Multiresolution Methods in Scattered Data Modelling.*

38. S.-I. Niculescu, K. Gu (eds.), *Advances in Time-Delay Systems.*

39. S. Attinger, P. Koumoutsakos (eds.), *Multiscale Modelling and Simulation.*

40. R. Kornhuber, R. Hoppe, J. Périaux, O. Pironneau, O. Wildlund, J. Xu (eds.), *Domain Decomposition Methods in Science and Engineering.*

41. T. Plewa, T. Linde, V.G. Weirs (eds.), *Adaptive Mesh Refinement – Theory and Applications.*

42. A. Schmidt, K.G. Siebert, *Design of Adaptive Finite Element Software.* The Finite Element Toolbox ALBERTA.

43. M. Griebel, M.A. Schweitzer (eds.), *Meshfree Methods for Partial Differential Equations II.*

44. B. Engquist, P. Lötstedt, O. Runborg (eds.), *Multiscale Methods in Science and Engineering.*

45. P. Benner, V. Mehrmann, D.C. Sorensen (eds.), *Dimension Reduction of Large-Scale Systems.*

46. D. Kressner, *Numerical Methods for General and Structured Eigenvalue Problems.*

47. A. Boriçi, A. Frommer, B. Joó, A. Kennedy, B. Pendleton (eds.), *QCD and Numerical Analysis III.*

48. F. Graziani (ed.), *Computational Methods in Transport.*

49. B. Leimkuhler, C. Chipot, R. Elber, A. Laaksonen, A. Mark, T. Schlick, C. Schütte, R. Skeel (eds.), *New Algorithms for Macromolecular Simulation.*

50. M. Bücker, G. Corliss, P. Hovland, U. Naumann, B. Norris (eds.), *Automatic Differentiation: Applications, Theory, and Implementations.*

51. A.M. Bruaset, A. Tveito (eds.), *Numerical Solution of Partial Differential Equations on Parallel Computers.*

52. K.H. Hoffmann, A. Meyer (eds.), *Parallel Algorithms and Cluster Computing.*

53. H.-J. Bungartz, M. Schäfer (eds.), *Fluid-Structure Interaction.*

54. J. Behrens, *Adaptive Atmospheric Modeling.*

55. O. Widlund, D. Keyes (eds.), *Domain Decomposition Methods in Science and Engineering XVI.*

56. S. Kassinos, C. Langer, G. Iaccarino, P. Moin (eds.), *Complex Effects in Large Eddy Simulations.*

57. M. Griebel, M.A Schweitzer (eds.), *Meshfree Methods for Partial Differential Equations III.*

58. A.N. Gorban, B. Kégl, D.C. Wunsch, A. Zinovyev (eds.), *Principal Manifolds for Data Visualization and Dimension Reduction.*

59. H. Ammari (ed.), *Modeling and Computations in Electromagnetics: A Volume Dedicated to Jean-Claude Nédélec.*

60. U. Langer, M. Discacciati, D. Keyes, O. Widlund, W. Zulehner (eds.), *Domain Decomposition Methods in Science and Engineering XVII.*

61. T. Mathew, *Domain Decomposition Methods for the Numerical Solution of Partial Differential Equations.*

62. F. Graziani (ed.), *Computational Methods in Transport: Verification and Validation.*

63. M. Bebendorf, *Hierarchical Matrices.* A Means to Efficiently Solve Elliptic Boundary Value Problems.

64. C.H. Bischof, H.M. Bücker, P. Hovland, U. Naumann, J. Utke (eds.), *Advances in Automatic Differentiation.*

65. M. Griebel, M.A. Schweitzer (eds.), *Meshfree Methods for Partial Differential Equations IV*.

66. B. Engquist, P. Lötstedt, O. Runborg (eds.), *Multiscale Modeling and Simulation in Science.*

67. I.H. Tuncer, Ü. Gülcat, D.R. Emerson, K. Matsuno (eds.), *Parallel Computational Fluid Dynamics 2007.*

68. S. Yip, T. Diaz de la Rubia (eds.), *Scientific Modeling and Simulations*.

69. A. Hegarty, N. Kopteva, E. O'Riordan, M. Stynes (eds.), *BAIL* 2008 – *Boundary and Interior Layers.*

70. M. Bercovier, M.J. Gander, R. Kornhuber, O. Widlund (eds.), *Domain Decomposition Methods in Science and Engineering XVIII.*

71. B. Koren, C. Vuik (eds.), *Advanced Computational Methods in Science and Engineering.*

72. M. Peters (ed.), *Computational Fluid Dynamics for Sport Simulation.*

73. H.-J. Bungartz, M. Mehl, M. Schäfer (eds.), *Fluid Structure Interaction II - Modelling, Simulation, Optimization.*

74. D. Tromeur-Dervout, G. Brenner, D.R. Emerson, J. Erhel (eds.), *Parallel Computational Fluid Dynamics 2008.*

75. A.N. Gorban, D. Roose (eds.), *Coping with Complexity: Model Reduction and Data Analysis.*

76. J.S. Hesthaven, E.M. Rønquist (eds.), *Spectral and High Order Methods for Partial Differential Equations.*

77. M. Holtz, *Sparse Grid Quadrature in High Dimensions with Applications in Finance and Insurance*.

78. Y. Huang, R. Kornhuber, O.Widlund, J. Xu (eds.), *Domain Decomposition Methods in Science and Engineering XIX*.

79. M. Griebel, M.A. Schweitzer (eds.), *Meshfree Methods for Partial Differential Equations V*.

80. P.H. Lauritzen, C. Jablonowski, M.A. Taylor, R.D. Nair (eds.), *Numerical Techniques for Global Atmospheric Models*.

81. C. Clavero, J.L. Gracia, F.J. Lisbona (eds.), *BAIL 2010 – Boundary and Interior Layers, Computational and Asymptotic Methods*.

82. B. Engquist, O. Runborg, Y.R. Tsai (eds.), *Numerical Analysis and Multiscale Computations*.

83. I.G. Graham, T.Y. Hou, O. Lakkis, R. Scheichl (eds.), *Numerical Analysis of Multiscale Problems*.

84. A. Logg, K.-A. Mardal, G. Wells (eds.), *Automated Solution of Differential Equations by the Finite Element Method.*

85. J. Blowey, M. Jensen (eds.), *Frontiers in Numerical Analysis - Durham 2010*.

86. O. Kolditz, U.-J. Gorke, H. Shao, W. Wang (eds.), *Thermo-Hydro-Mechanical-Chemical Processes in Fractured Porous Media - Benchmarks and Examples*.

87. S. Forth, P. Hovland, E. Phipps, J. Utke, A. Walther (eds.), *Recent Advances in Algorithmic Differentiation*.

88. J. Garcke, M. Griebel (eds.), *Sparse Grids and Applications*.

89. M. Griebel, M.A. Schweitzer (eds.), *Meshfree Methods for Partial Differential Equations VI.*

90. C. Pechstein, *Finite and Boundary Element Tearing and Interconnecting Solvers for Multiscale Problems*.

91. R. Bank, M. Holst, O. Widlund, J. Xu (eds.), *Domain Decomposition Methods in Science and Engineering XX*.

92. H. Bijl, D. Lucor, S. Mishra, C. Schwab (eds.), *Uncertainty Quantification in Computational Fluid Dynamics*.

93. M. Bader, H.-J. Bungartz, T. Weinzierl (eds.), *Advanced Computing*.

94. M. Ehrhardt, T. Koprucki (eds.), *Advanced Mathematical Models and Numerical Techniques for Multi-Band Effective Mass Approximations*.

95. M. Azaïez, H. El Fekih, J.S. Hesthaven (eds.), *Spectral and High Order Methods for Partial Differential Equations ICOSAHOM 2012*.

96. F. Graziani, M.P. Desjarlais, R. Redmer, S.B. Trickey (eds.), *Frontiers and Challenges in Warm Dense Matter*.

97. J. Garcke, D. Pflüger (eds.), *Sparse Grids and Applications – Munich 2012*.

98. J. Erhel, M. Gander, L. Halpern, G. Pichot, T. Sassi, O. Widlund (eds.), *Domain Decomposition Methods in Science and Engineering XXI.*

99. R. Abgrall, H. Beaugendre, P.M. Congedo, C. Dobrzynski, V. Perrier, M. Ricchiuto (eds.), *High Order Nonlinear Numerical Methods for Evolutionary PDEs - HONOM 2013.*

100. M. Griebel, M.A. Schweitzer (eds.), *Meshfree Methods for Partial Differential Equations VII.*

101. R. Hoppe (ed.), *Optimization with PDE Constraints - OPTPDE 2014*.

102. S. Dahlke, W. Dahmen, M. Griebel, W. Hackbusch, K. Ritter, R. Schneider, C. Schwab, H. Yserentant (eds.), *Extraction of Quantifiable Information from Complex Systems*.

103. A. Abdulle, S. Deparis, D. Kressner, F. Nobile, M. Picasso (eds.), *Numerical Mathematics and Advanced Applications - ENUMATH 2013*.

104. T. Dickopf, M.J. Gander, L. Halpern, R. Krause, L.F. Pavarino (eds.), *Domain Decomposition Methods in Science and Engineering XXII*.

105. M. Mehl, M. Bischoff, M. Schäfer (eds.), *Recent Trends in Computational Engineering - CE2014*. Optimization, Uncertainty, Parallel Algorithms, Coupled and Complex Problems.

106. R.M. Kirby, M. Berzins, J.S. Hesthaven (eds.), *Spectral and High Order Methods for Partial Differential Equations - ICOSAHOM'14*.

107. B. Jüttler, B. Simeon (eds.), *Isogeometric Analysis and Applications 2014*.

108. P. Knobloch (ed.), *Boundary and Interior Layers, Computational and Asymptotic Methods – BAIL 2014*.

109. J. Garcke, D. Pflüger (eds.), *Sparse Grids and Applications – Stuttgart 2014*.

110. H. P. Langtangen, *Finite Difference Computing with Exponential Decay Models*.

111. A. Tveito, G.T. Lines, *Computing Characterizations of Drugs for Ion Channels and Receptors Using Markov Models*.

112. B. Karasözen, M. Manguoğlu, M. Tezer-Sezgin, S. Göktepe, Ö. Uğur (eds.), *Numerical Mathematics and Advanced Applications ENUMATH 2015*.

113. H.-J. Bungartz, P. Neumann, W. Nagel (eds.), *Software for Exascale Computing - SPPEXA 2013-2015*.

114. G.R. Barrenechea, F. Brezzi, A. Cangiani, E.H. Georgoulis (eds.), *Building Bridges: Connections and Challenges in Modern Approaches to Numerical Partial Differential Equations*.

For further information on these books please have a look at our mathematics catalogue at the following URL: www.springer.com/series/3527

Monographs in Computational Science and Engineering

1. J. Sundnes, G.T. Lines, X. Cai, B.F. Nielsen, K.-A. Mardal, A. Tveito, *Computing the Electrical Activity in the Heart.*

For further information on this book, please have a look at our mathematics catalogue at the following URL: www.springer.com/series/7417

Texts in Computational Science and Engineering

1. H. P. Langtangen, *Computational Partial Differential Equations.* Numerical Methods and Diffpack Programming. 2nd Edition
2. A. Quarteroni, F. Saleri, P. Gervasio, *Scientific Computing with MATLAB and Octave.* 4th Edition
3. H. P. Langtangen, *Python Scripting for Computational Science*. 3rd Edition
4. H. Gardner, G. Manduchi, *Design Patterns for e-Science*.
5. M. Griebel, S. Knapek, G. Zumbusch, *Numerical Simulation in Molecular Dynamics*.
6. H. P. Langtangen, *A Primer on Scientific Programming with Python.* 5th Edition
7. A. Tveito, H. P. Langtangen, B. F. Nielsen, X. Cai, *Elements of Scientific Computing.*
8. B. Gustafsson, *Fundamentals of Scientific Computing.*
9. M. Bader, *Space-Filling Curves.*
10. M. Larson, F. Bengzon, *The Finite Element Method: Theory, Implementation and Applications.*
11. W. Gander, M. Gander, F. Kwok, *Scientific Computing: An Introduction using Maple and MATLAB.*
12. P. Deuflhard, S. Röblitz, *A Guide to Numerical Modelling in Systems Biology*.
13. M. H. Holmes, *Introduction to Scientific Computing and Data Analysis*.
14. S. Linge, H. P. Langtangen, *Programming for Computations* - A Gentle Introduction to Numerical Simulations with MATLAB/Octave.
15. S. Linge, H. P. Langtangen, *Programming for Computations* - A Gentle Introduction to Numerical Simulations with Python.

For further information on these books please have a look at our mathematics catalogue at the following URL: www.springer.com/series/5151