Andrea Cangiani
Ruslan L. Davidchack
Emmanuil Georgoulis
Alexander N. Gorban
Jeremy Levesley
Michael V. Tretyakov *Editors*

Numerical Mathematics and Advanced Applications

ENUMATH 2011

Springer

Numerical Mathematics and Advanced Applications 2011

Andrea Cangiani • Ruslan L. Davidchack
Emmanuil Georgoulis • Alexander N. Gorban
Jeremy Levesley • Michael V. Tretyakov
Editors

Numerical Mathematics and Advanced Applications 2011

Proceedings of ENUMATH 2011,
the 9th European Conference on Numerical
Mathematics and Advanced Applications,
Leicester, September 2011

Springer

Editors
Andrea Cangiani
Ruslan L. Davidchack
Emmanuil Georgoulis
Alexander N. Gorban
Jeremy Levesley
Department of Mathematics
University of Leicester
Leicester
United Kingdom

Michael V. Tretyakov
School of Mathematical Sciences
University of Nottingham
Nottingham
United Kingdom

ISBN 978-3-642-33133-6 ISBN 978-3-642-33134-3 (eBook)
DOI 10.1007/978-3-642-33134-3
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012955743

Math. Subj. Class. (2010): 65-06, 65M60, 65M15, 65M06, 35Q35, 35K57, 35Q60

Preface

The European Conference on Numerical Mathematics and Advanced Applications (ENUMATH) is an established series of conferences held every 2 years to provide a forum for discussion on recent aspects of numerical mathematics and challenging scientific and industrial applications at the highest level of international expertise. The ENUMATH conferences previously took place in Paris (1995), Heidelberg (1997), Jyvaskyla (1999), Ischia (2001), Prague (2003), Santiago de Compostela (2005), Graz (2007), Uppsala (2009). This volume contains a selection of papers presented at ENUMATH 2011, organised by the Department of Mathematics, University of Leicester, UK, and held at Leicester's Main Campus between September 5–9, 2011.

The 2011 edition of ENUMATH attracted about 300 participants from around the world, including ten invited talks by:

- J.-F. Gerbeau (Inria-Rocquenc., France), on "Direct and inverse modeling in hemodynamics";
- V. Girault (Paris, France), "On the coupling of Stokes or Navier-Stokes and Darcy flows through porous media";
- I. Graham (Bath, UK), on the "Solution of elliptic PDEs with high contrast heterogeneous coefficients ";
- T. Lelievre (Cermics/U. Paris 6, France), on "Sampling techniques in molecular dynamics";
- V. Simoncini (Bologna, Italy), on "Iterative solvers for saddle point algebraic linear systems: tools of the trade";
- C.-W. Shu (Brown, USA), on "Maximum-principle-satisfying and positivity-preserving high order discontinuous Galerkin and finite volume schemes";
- A. Stuart (Warwick, UK), on "Filtering the Navier-Stokes Equation";
- S. Turek (Dortmund, Germany), on "Hardware-oriented Numerics (for PDE) - Motivation, Concepts, Software";
- K. Urban (Ulm, Germany), on "Reduced Basis Methods for Optimization in Industrial Challenges", and
- R. Winther (Oslo, Norway), on "Bounded cochain projections, why and how",

28 minisymposia on various aspects of numerical and applied mathematics, a large number of contributed talks and a public lecture by Prof. Nicholas J. Higham, FRS (Manchester, UK), on "Numerical Linear Algebra in the UK: from Cayley to Exascale Computing." The conference made a contribution to development of numerical mathematics internationally and, in particular, in the UK.

We hope that this ENUMATH 2011 Proceedings Volume will appeal to a wide range of readers, giving an overview and recent developments in computational mathematics, their applications and some related fields. A total of 87 contributions appear in the proceedings of ENUMATH 2011, from a wide range of topics: from theory and analysis of numerical methods, applications in biology, finance, physics and engineering, to high performance algorithms for scientific computing. The volume is organised in 11 parts, namely:

I A Posteriori Error Estimation and Adaptive Methods;
II Computational Electromagnetics;
III Computational Methods;
IV Convection-diffusion, Conservation laws, and Hyperbolic Systems;
V Discontinuous Galerkin Methods;
VI Finite Element and Finite Volume techniques;
VII Fluid Mechanics;
VIII High Performance Computing;
IX Multiscale Modeling and Simulations;
X Preconditioners and Solvers;
XI Uncertainty, Stochastic Modelling, and Applications.

A number of contributions could be attributed to more than one part in terms of their content. In these cases the criterion of the coherence of the proceedings was also taken into consideration.

We would like to thank all the contributors for submitting their work for inclusion in the Proceedings Volume and for their timely response during the reviewing process. We express our sincere gratitude to all the participants of ENUMATH 2011 for their attendance, valuable scientific contributions, and stimulating discussions during the conference. We particularly extend our thanks to the minisymposia organisers who did an excellent job in putting them together coherently and within the given time restrictions, and to the chairs of the contributed sessions for agreeing to serve as such. A big share of the success of this conference series has been due to the members of the Programme Committee: F. Brezzi, M. Feistauer, R. Glowinski, R. Jeltsch, Y. Kuznetsov, J. Periaux, and R. Rannacher. We would also like to thank the Scientific Committee of ENUMATH 2011 for agreeing to give valuable input into this Proceedings Volume and for their general support of the conference.

Special gratitude goes to our sponsors: the Institute of Mathematics and its Applications, DataVisibility, Associated Architects, Oxford University Press, Numerical Algorithms Group (NAG), the Department of Mathematics at Leicester, and Leicester Conferences for supporting ENUMATH 2011 in a number of ways. We would also like to thank Springer-Verlag for their support and patience while preparing this volume.

This conference would not have been made possible without the tireless efforts of the staff and the PhD students of the Department of Mathematics at Leicester, to whom a big thanks is due. Very special thanks goes to our Administrator Tara Chakraborti for taking excellent care of the bulk of organisation of the conference, and to Dan Carter for doing a great job at keeping everything running smoothly.

We believe that Leicester ENUMATH 2011 was a very interesting and worthwhile experience, and we hope that the participants had a good and fruitful time in Leicester.

Leicester, UK

The Local Organising Committee:

Andrea Cangiani
Ruslan Davidchack
Manolis Georgoulis
Alexander Gorban
Jeremy Levesley
Michael Tretyakov

Contents

An Adaptive MFD Method for the Obstacle Problem

P.F. Antonietti, L. Beirão da Veiga, and M. Verani

Abstract We present an adaptive mimetic finite difference method for the approximate solution of variational inequalities. The adaptive strategy is based on a heuristic hierarchical type error indicator. Numerical experiments that validate the performance of the adaptive MFD method are also presented.

1 The Obstacle Problem

Throughout the paper we will use standard notations for Sobolev spaces, norms and seminorms. For a bounded domain D in $\mathbb{R}^2$, we denote by $H^s(D)$ the standard Sobolev space of order $s \geq 0$, and by $\|\cdot\|_{H^s(D)}$ and $|\cdot|_{H^s(D)}$ the usual Sobolev norm and seminorm, respectively. For $s = 0$, we write $L^2(D)$ instead of $H^0(D)$. $H^1_0(D)$ is the subspace of $H^1(D)$ of functions with zero trace on ∂D.

Let Ω be an open, bounded, convex set of $\mathbb{R}^2$, with either a polygonal or a C^2-smooth boundary $\Gamma := \partial\Omega$. Let $g := \tilde{g}_{|\Gamma}$, with $\tilde{g} \in H^2(\Omega)$ and we set $V^g := \{v \in H^1(\Omega) : \ v = g \text{ on } \Gamma\}$. Let us introduce the bilinear form $a(\cdot,\cdot) : H^1(\Omega) \times H^1(\Omega) \longrightarrow \mathbb{R}$ defined by $a(u,v) := \int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x$, and the linear functional $F(\cdot) : H^1(\Omega) \longrightarrow \mathbb{R}$ with $F(v) := \int_\Omega f \, v \, \mathrm{d}x$, where we assume $f \in L^2(\Omega)$. Finally, we define the function $\psi \in H^2(\Omega)$ with $\psi \leq g$ on Γ and the convex space

$$K := \{v \in V^g \ : \ v \geq \psi \text{ a.e. in } \Omega\}.$$

P.F. Antonietti · M. Verani (✉)
MOX – Modelling and Scientific Computing - Dipartimento di Matematica "F. Brioschi", Politecnico di Milano, Piazza Leonardo da Vinci 32, I-20133, Milano, Italy
e-mail: paola.antonietti@polimi.it; marco.verani@polimi.it

L. Beirão da Veiga
Dipartimento di Matematica, Università di Milano, Via Saldini 50, I-20133, Milano, Italy
e-mail: lourenco.beirao@unimi.it

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_1,

We are interested in solving the following variational inequality:

$$\begin{cases} \text{Find } u \in K \text{ such that} \\ a(u, v-u) \geq F(v-u) \qquad \forall v \in K. \end{cases} \tag{1}$$

It is well known (see e.g. [6]) that under the above data regularity assumption, the elliptic obstacle problem (1) admits a unique solution $u \in H^2(\Omega)$.

2 The Mimetic Discretization

In this section we recall the mimetic discretization for the obstacle problem (1) (see [2] for more details). Let $\Omega_h \subset \Omega$ be a polygonal approximation of Ω, in such a way that all vertexes of Ω_h which are on the boundary of Ω_h are also on the boundary of Ω. The polygonal domain Ω_h represents the computational domain for the method. With a little abuse of notation, we also denote by Ω_h a partition of the above introduced computational domain into polygons E. We assume that this partition is conformal, i.e., intersection of two different elements E_1 and E_2 is either a few mesh points, or a few mesh edges (two adjacent elements may share more than one edge) or empty. We allow Ω_h to contain non-convex elements. Note moreover that, differently from conforming finite element meshes, T-junctions are now allowed in the mesh; indeed, this are included in the above conditions simply by splitting single edges into two new (aligned) edges. For each polygon E, k_E denotes its number of vertexes, $|E|$ its area, h_E its diameter and

$$h := \max_{E \in \Omega_h} h_E.$$

We denote the set of mesh vertexes and edges by $\mathscr{N}_h$ and $\mathscr{E}_h$, the set of internal vertexes and edges by $\mathscr{N}_h^0$ and $\mathscr{E}_h^0$, the set of boundary vertexes and edges by $\mathscr{N}_h^\partial$ and $\mathscr{E}_h^\partial$. The set of vertexes and edges of a particular element E are denoted by $\mathscr{N}_h^E$ and $\mathscr{E}_h^E$, respectively. Moreover, we denote a generic mesh vertex by v, a generic edge by e and its length both by h_e and $|\mathsf{e}|$. A fixed orientation is also set for the mesh Ω_h, which is reflected by a unit normal vector $\mathbf{n}_\mathsf{e}$, $\mathsf{e} \in \mathscr{E}_h$, fixed once for all. For every polygon E and edge $\mathsf{e} \in \mathscr{E}_h^E$, we define a unit normal vector $\mathbf{n}_E^\mathsf{e}$ that points outside E.

The mesh is assumed to satisfy the following shape regularity properties, which have already been used in [7]. There exist

- An integer number N_s independent of h;
- A real positive number ρ independent of h;
- A *compatible* sub-decomposition $\mathscr{T}_h$ of every Ω_h into shape-regular triangles,

such that

(H1) Any polygon $E \in \Omega_h$ admits a decomposition $\mathscr{T}_h|_E$ formed by less than N_s triangles;
(H2) Any triangle $T \in \mathscr{T}_h$ is shape-regular in the sense that the ratio between the radius r_T of the inscribed ball and the diameter h_T of T is bounded from below by ρ; i.e. $0 < \rho \leq \frac{r_T}{h_T}$.

The discretization of problem (1) requires to discretize a scalar field in $H^1(\Omega)$. To this aim, we start introducing the degrees of freedom for the discrete approximation space. The discrete space V_h is defined as follows: a vector $v_h \in V_h$ consists of a collection of degrees of freedom

$$v_h := \{v^{\mathsf{v}}\}_{\mathsf{v} \in \mathscr{N}_h},$$

one per mesh vertex, e.g. to every vertex $\mathsf{v} \in \mathscr{N}_h$, we associate a real number v^{v}. The scalar v^{v} represents the nodal value of the underlying discrete scalar field. The number of unknowns is equal to the number of vertexes of the mesh. We also define the discrete space $V_h^g \subset V_h$ of functions which satisfy the Dirichlet boundary condition:

$$V_h^g := \{v_h \in V_h \ : \ v_h^{\mathsf{v}} = g(\mathsf{v}) \ \forall \mathsf{v} \in \mathscr{N}_h^{\partial}\}.$$

Accordingly, V_h^0 represents the space of discrete functions which vanish at the boundary nodes.

We define the following interpolation operator from the spaces of smooth enough functions to the discrete space V_h. For every function $v \in \mathscr{C}^0(\bar{\Omega}) \cap H^1(\Omega)$, we define $v_{\mathrm{I}} \in V_h$ by

$$v_{\mathrm{I}}^{\mathsf{v}} := v(\mathsf{v}) \quad \forall \mathsf{v} \in \mathscr{N}_h.$$

Moreover, we analogously define the local interpolation operator from $\mathscr{C}^0(\bar{E}) \cap H^1(E)$ into $V_h|_E$ given by

$$v_{\mathrm{I}}^{\mathsf{v}} := v(\mathsf{v}) \quad \forall \mathsf{v} \in \mathscr{N}_h^E.$$

We endow the space V_h with the following discrete seminorm

$$\|v_h\|_{1,h}^2 := \sum_{E \in \Omega_h} \|v_h\|_{1,h,E}^2 = \sum_{E \in \Omega_h} |E| \sum_{\mathsf{e} \in \mathscr{E}_h^E} \left[\frac{1}{|\mathsf{e}|} (v^{\mathsf{v}_2} - v^{\mathsf{v}_1}) \right]^2, \tag{2}$$

where v_1 and v_2 are the two vertexes of e. The quantity $\|\cdot\|_{1,h}$ is a $H^1(\Omega)$-type discrete seminorm, which becomes a norm on V_h^0. We denote by $a_h(\cdot,\cdot) : V_h \times V_h \to \mathbb{R}$ the discretization of the bilinear form $a(\cdot,\cdot)$, defined as follows:

$$a_h(v_h, w_h) := \sum_{E \in \Omega_h} a_h^E(v_h, w_h) \quad \forall v_h, w_h \in V_h, \tag{3}$$

where $a_h^E(\cdot,\cdot)$ is a symmetric bilinear form on each element E. We introduce two fundamental assumptions for the local bilinear form $a_h^E(\cdot,\cdot)$. The first one represents the coercivity (up to the kernel) and the correct scaling with respect to the element size.

(S1) There exist two positive constants c_1 and c_2 independent of h such that, for every $u_h, v_h \in V_h$ and each $E \in \Omega_h$, we have

$$c_1\|v_h\|_{1,h,E}^2 \le a_h^E(v_h,v_h), \qquad a_h^E(u_h,v_h) \le c_2\|u_h\|_{1,h,E}\|v_h\|_{1,h,E}.$$

(S2) For every element E, every linear vector function p^1 on E, and every $v_h \in V_h$, it holds

$$a_h^E(v_h,(p^1)_1) = \sum_{\mathsf{e}\in\mathscr{E}_h^E} (\nabla p^1\cdot\mathbf{n}_E^{\mathsf{e}})\frac{|\mathsf{e}|}{2}\big(v_h^{\mathsf{v}_1}+v_h^{\mathsf{v}_2}\big), \tag{4}$$

where v_1 and v_2 are the two vertexes of $\mathsf{e}\in\mathbf{n}_E^{\mathsf{e}}$.

We remark that the meaning of the consistency condition (S2) is that the discrete bilinear form respects integration by parts when tested with linear functions. The bilinear form $a_h(\cdot,\cdot)$ can be easily built element by element in a simple algebraic way; see for instance [2, 7]. Finally, we are able to define the proposed mimetic discrete method for the obstacle problem. Let the loading term

$$(f,v_h)_h := \sum_{E\in\Omega_h} \bar{f}|_E \sum_{i=1}^{k_E} v^{\mathsf{v}_i}\omega_E^i, \tag{5}$$

where $\mathsf{v}_1,\dots,\mathsf{v}_{k_E}$ are the vertexes of E, $\bar{f}|_E := \frac{1}{|E|}\int_E f\,\mathrm{d}x$, and $\omega_E^1,\dots,\omega_E^{k_E}$ are positive weights such that $\sum_{i=1}^{k_E}\omega_E^i = |E|$. Finally, let us introduce the discrete convex space

$$K_h := \{v_h\in V_h^g \ :\ v_h^{\mathsf{v}} \ge \psi(\mathsf{v})\ \forall \mathsf{v}\in\mathscr{N}_h\}.$$

Then, the mimetic discretization of problem (1) reads:

$$\begin{cases} \text{Find } u_h\in K_h \text{ such that} \\ a_h(u_h,v_h-u_h) \ge (f,v_h-u_h)_h \qquad \forall v_h\in K_h. \end{cases} \tag{6}$$

Thanks to property (S1) it is immediate to check that the bilinear form $a_h(\cdot,\cdot)$ is coercive on $V_h/\mathbb{R}$. As a consequence, recalling again that $K_h\subset V_h$ is convex and closed, standard results [8] give the existence and uniqueness of a solution for the discrete problem (6). The following convergence result has been proved in [2].

Theorem 1. *Let $u\in K\cap H^2(\Omega)$ be the solution to the continuous problem (1), and $u_h\in K_h$ be the corresponding mimetic approximation, obtained by solving the discrete problem (6). Then, it holds*

$$\|u_h - u_I\|_{1,h} \leq C\,h,$$

where the constant C is independent of the mesh-size h.

3 An Adaptive MFD Algorithm

In this section we extend the h-adaptive MFD algorithm presented in [3] to the case of the obstacle problem (1). The adaptive procedure, based on a posteriori error indicator of hierarchical type, has the following form:

$$\texttt{SOLVE} \rightarrow \texttt{ESTIMATE} \rightarrow \texttt{MARK} \rightarrow \texttt{REFINE}.$$

Here `SOLVE` computes the discrete solution to (6). The module `ESTIMATE` makes use of a suitable fluctuation problem (cf. (12) below) to build the hierarchical error indicators, while the procedure `MARK` employs the fixed fraction strategy, with refinement fraction set to 30 %, to make a selection of the elements to be refined. Finally, the module `REFINE` uses the strategy described in Sect. 3.1 to subdivide elements marked for refinement. In the next two sections we will briefly describe the modules `REFINE` and `ESTIMATE`.

3.1 Mesh Refinement

Given a mesh Ω_h we can build a uniformly refined mesh $\widehat{\Omega}_h$ as follows. We start assuming that

(H3) All polygons $E \in \Omega_h$ are convex.

Then, we introduce the point $\mathbf{x}_E \in E$

$$\mathbf{x}_E := \frac{1}{N} \sum_{\mathsf{v} \in \partial E} \mathbf{x}(\mathsf{v})\,, \tag{7}$$

where N is the number of vertexes in ∂E and $\mathbf{x}(\mathsf{v})$ is the position vector of node $\mathsf{v} \in \mathcal{N}$.

Remark 1. We remark that assumption (H3) is made essentially for the sake of exposition. What follows can be adapted to cover more general cases such as, for instance, elements which are star shaped with respect to a ball. In particular, (7) has to be modified to define an interior point, and (10) has to be changed, for $\mathsf{v} = \mathbf{x}_E$, in such a way that the operator preserves the linear functions.

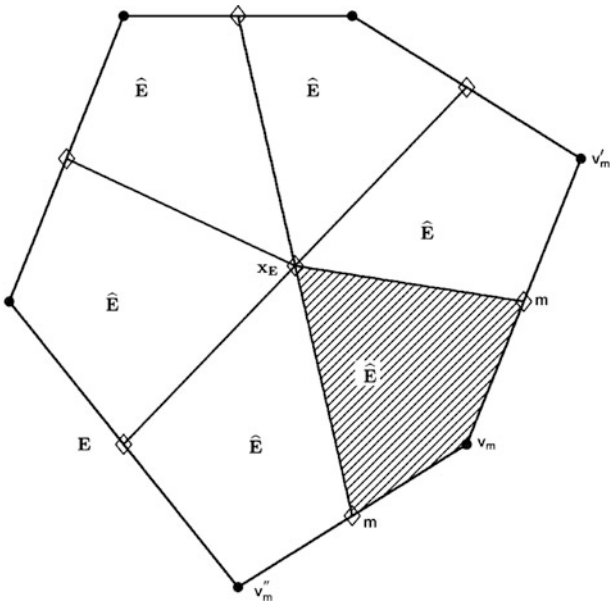


Fig. 1 Refinement strategy: coarse element $E \in \Omega_h$ and sub-elements $\widehat{E} \in \widehat{\Omega}_h$. *Circles* denote the coarse vertexes, while diamonds refer to additional vertexes in the finer mesh

The uniformly refined mesh $\widehat{\Omega}_h$ is built by subdividing each element E of Ω_h in the following way: each midpoint $\mathsf{m} = \mathsf{m}(e)$ of each edge $e \in \partial E$ is connected with the point $\mathbf{x}_E$. This determines a subdivision of E into sub-elements which are collected for all $E \in \Omega_h$ to form the new mesh $\widehat{\Omega}_h$ (see Fig. 1). In the following, we will indicate all geometrical objects of the finer grid $\widehat{\Omega}_h$ with a hat symbol, the meaning being the same as in the original mesh. For instance, we will indicate with $\widehat{E}$ a generic element of $\widehat{\Omega}_h$, and with $\widehat{\mathcal{N}_h}$ the set of all its vertexes. Note that

$$\widehat{\mathcal{N}_h} = \mathcal{N}_h \cup \{\mathsf{m}(e)\}_{e\in\mathscr{E}} \cup \{\mathbf{x}_E\}_{E\in\Omega_h},$$

i.e. the edge midpoints $\mathsf{m}(e)$ and the points $\mathbf{x}_E$ become additional vertexes in the new mesh $\widehat{\Omega}_h$. In addition, $\widehat{h}$ will denote the mesh-size of the finer mesh $\widehat{\Omega}_h$, i.e. $\widehat{h} = \max_{\widehat{E}\in\widehat{\Omega}_h} h_{\widehat{E}}$.

3.2 Hierachical Error Indicators

Following the construction given in Sect. 2, we can introduce the finer discrete spaces $\widehat{V}_h$ and $\widehat{K}_h$ associated to the mesh $\widehat{\Omega}_h$, a bilinear form $\widehat{a}_h(\cdot,\cdot) : \widehat{V}_h \times \widehat{V}_h \to \mathbb{R}$ and a suitable loading term, so that the finer version of the coarse discrete

problem (6) reads as follows

$$\begin{cases} \text{Find } \widehat{u}_h \in \widehat{K}_h \text{ such that} \\ \widehat{a}_h(\widehat{u}_h, v_h - \widehat{u}_h) \geq (f, v_h - \widehat{u}_h)_{\widehat{h}} \qquad \forall v_h \in \widehat{K}_h. \end{cases} \tag{8}$$

We now introduce two operators that maps the finer space into the coarser one and viceversa. Let $\Pi : \widehat{V}_h \to V_h$ be defined by

$$\big(\Pi(v_h)\big)(\mathsf{v}) = v_h(\mathsf{v}) \quad \forall \mathsf{v} \in \mathscr{N}_h, \ \forall v_h \in \widehat{V}_h. \tag{9}$$

Given any midpoint $\mathsf{m} = \mathsf{m}(e)$, $e \in \mathscr{E}_h$, we indicate with v_m and with v'_m the two vertexes which are endpoints to the edge e. We then define $\Pi^\dagger : V_h \to \widehat{V}_h$ by

$$\big(\Pi^\dagger(v_h)\big)(\mathsf{v}) = \begin{cases} v_h(\mathsf{v}) & \forall \mathsf{v} \in \mathscr{N}_h \\ \dfrac{1}{2}\big(v_h(\mathsf{v}_\mathsf{m}) + v_h(\mathsf{v}'_\mathsf{m})\big) & \text{if } \mathsf{v} = \mathsf{m}(e), \ e \in \mathscr{E} \\ \dfrac{1}{N}\displaystyle\sum_{\mathsf{v} \in \partial E} v_h(\mathsf{v}) & \text{if } \mathsf{v} = \mathbf{x}_E, \ E \in \Omega_h, \end{cases} \tag{10}$$

for all $v_h \in V_h$. The operator $\Pi^\dagger$ embeds the coarse space V_h into the finer space $\widehat{V}_h$ by averaging the coarse vertex values. We denote by $\widehat{V}_h^c$ the subspace of $\widehat{V}_h$ given by the image of $\Pi^\dagger$ and we refer to it as to the embedded coarse space. Finally, we introduce the fluctuation space

$$\widehat{V}_h^f = \{v_h \in \widehat{V}_h \mid v_h(\mathsf{v}) = 0 \ \forall \mathsf{v} \in \mathscr{N}_h\}.$$

It is immediate to check that

$$\widehat{V}_h = \widehat{V}_h^c \oplus \widehat{V}_h^f.$$

Let $\|\cdot\|_{1,\widehat{h}}$ and $\|\cdot\|_{1,\widehat{h},\widehat{E}}$, $\widehat{E} \in \widehat{V}_h$, denote the global and local norms of the finer space $\widehat{V}_h$ (cf. (2)). Accordingly, we indicate with $\|\cdot\|_{1,\widehat{h},E}$ the norm of the fine space restricted to the coarse element $E \in \Omega_h$

$$\|v_h\|^2_{1,\widehat{h},E} = \sum_{\widehat{E} \in E} \|v_h\|^2_{1,\widehat{h},\widehat{E}} \quad \forall v_h \in \widehat{V}_h.$$

For all $E \in \Omega_h$, we can define a bilinear form $a_h^E(\cdot,\cdot)$ on the coarse space V_h as follows

$$a_h^E(v_h, w_h) := \sum_{\widehat{E} \in E} \widehat{a}_h^{\widehat{E}}(\Pi^\dagger(v_h), \Pi^\dagger(w_h)) \quad \forall v_h, w_h \in V_h \ . \tag{11}$$

Similarly, we can also define a loading term

$$(f, v_h)_h = \sum_{E \in \Omega_h} \sum_{\widehat{E} \in E} (f, \Pi^\dagger(v_h))_{\widehat{h},\widehat{E}}$$

where $(f, \cdot)_{\widehat{h},\widehat{E}}$ represents a local scalar product on the fine mesh constructed analogously to (5).

Inspired by Zou et al. [10] (see also [1, 4, 5]) we introduce the following fluctuation discrete problem:

$$\begin{cases} \text{Find } \widehat{e}_h^f \in \widehat{K}_h^f \text{ such that} \\ \widehat{a}_h(\widehat{e}_h^f, v_h^f - \widehat{e}_h^f) \geq (f, v_h^f - \widehat{e}_h^f)_{\widehat{h}} - \widehat{a}_h(\Pi^\dagger u_h, v_h^f - \widehat{e}_h^f) \quad \forall v_h^f \in \widehat{K}_h^f \,, \end{cases} \tag{12}$$

where

$$\widehat{K}_h^f = \{v_h^f \in \widehat{V}_h^f \,:\, v_h^f(\mathsf{v}) \geq \psi(\mathsf{v}) - \Pi^\dagger u_h(\mathsf{v}) \quad \forall \mathsf{v} \in \widehat{\mathcal{N}_h} \setminus \mathcal{N}_h\}.$$

Note that the right-hand side in (12) is the residual of the approximate solution u_h when tested with the fluctuation space $\widehat{K}_h^f$. The local heuristic error indicators are $\eta_E := \sum_{\widehat{E} \in E} \|\widehat{e}_h^f\|^2_{1,\widehat{h},\widehat{E}}$ being $\widehat{e}_h^f$ the solution of problem (12), while we set $\eta^2 = \sum_{E \in \Omega_h} \eta_E^2$. The quantities η_E, computed in the module `ESTIMATE`, are employed by the procedure `MARK` to select the elements of the mesh Ω_h to be refined. We refer to [3] (where a similar approach has been employed for the solution of linear elliptic problems) for more details on the construction of possible inexpensive heuristic variants of the error indicators η_E.

3.3 Numerical Results

Next we investigate the numerical performance of our adaptive MFD method. We consider the domain $\Omega =]-1, 1[^2$. For the parameter $r = 0.7$, we define the (continuous) load

$$f(x, y) := \begin{cases} -8(2x^2 + 2y^2 - r^2) & \text{if } \sqrt{x^2 + y^2} > r, \\ -8r^2(1 - x^2 - y^2 + r^2) & \text{if } \sqrt{x^2 + y^2} \leq r, \end{cases} \tag{13}$$

and the Dirichlet boundary data $g(x, y) := (x^2 + y^2 - r^2)^2$. We consider a constant obstacle $\psi(x, y) := 0$, so that the exact minimizer of model problem (1) is given by

$$u(x, y) := (\max\{x^2 + y^2 - r^2, 0\})^2; \tag{14}$$

cf. [9].

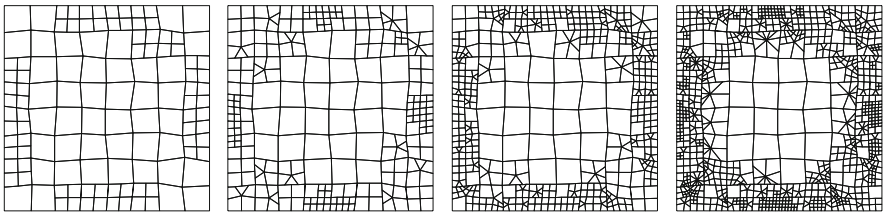


Fig. 2 First four levels of computational meshes generated by the adaptive refinement strategy employing the fixed fraction marking strategy

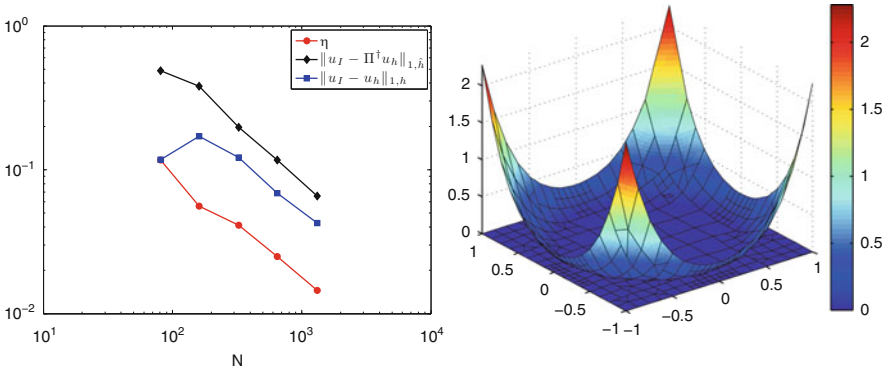


Fig. 3 *Left*: actual errors and error indicator versus the number of degrees of freedom (loglog-scale). The adaptive meshes are constructed by employing the fixed fraction marking strategy. *Right*: adaptive approximate solution after four iterations

In Fig. 2 we report the first four levels of meshes generated by the adaptive algorithm employing the fixed fraction marking strategy. We observe that the mesh is correctly refined along the boundary of the contact region and not in its interior where the solution (equal to the obstacle) is indeed smooth. In Fig. 3 the error estimator computed on the sequence of the adaptively generated meshes together with the actual error in the discrete energy norm and the error $\|u - \Pi^\dagger u_h\|_{1,\widehat{h}}$ are plotted as a function of the number of degrees of freedom (loglog-scale).

References

1. M. Ainsworth and J. T. Oden. *A posteriori error estimation in finite element analysis*. Pure and Applied Mathematics (New York). Wiley-Interscience [John Wiley & Sons], New York, 2000.
2. P. F. Antonietti, L. Beirão da Veiga, and M. Verani. A mimetic discretization of elliptic obstacle problems. To appear on Math. Comp.
3. P. F. Antonietti, L. Beirão da Veiga, C. Lovadina and M. Verani. Hierarchical a posteriori error estimators for the mimetic discretization of elliptic problems. Technical Report 33, MOX, Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci, 32, 20133 Milano, Italy, 2010. http://mox.polimi.it/it/progetti/pubblicazioni/.

4. R. E. Bank. Hierarchical bases and the finite element method. *Acta Numerica*, 5:1–43, 1996.
5. F. A. Bornemann, B. Erdmann, and R. Kornhuber. A posteriori error estimates for elliptic problems in two and three space dimensions. *SIAM J. Numer. Anal.*, 33(3):1188–1204, 1996.
6. H. Brezis and G. Stampacchia. Sur la régularité de la solution d'inéquations elliptiques. *Bull. Soc. Math. France*, 96:153–180, 1968.
7. F. Brezzi, A. Buffa, and K. Lipnikov. Mimetic finite differences for elliptic problems. *M2AN Math. Model. Numer. Anal.*, 43(2):277–295, 2009.
8. P. G. Ciarlet, *The finite element method for elliptic problems*, North-Holland, Amsterdam, 1978.
9. R. H. Nochetto, K. G. Siebert, and A. Veeser, Pointwise a posteriori error control for elliptic obstacle problems. *Numer. Math.*, 95(1):163–195, 2003
10. Q. Zou, A. Veeser, R. Kornhuber, C. Graser Hierarchical error estimates for the energy functional in obstacle problems. *Numer. Math.*, 117(4):653–677, 2011.

Reconstruction-Based a Posteriori Error Estimators for the Transport Equation

R. Becker, D. Capatina, and R. Luce

Abstract We present a unified approach to build error estimators based on $H(\mathrm{div})$-reconstructed fluxes on the primal mesh, inspired by the hypercircle method. Here, the transport equation is considered and discretized by discontinuous Galerkin, nonconforming and conforming finite elements. We describe the local computation of fluxes on patches, obtain upper error bounds and show some numerical tests.

1 Introduction

In order to achieve mesh adaptivity, one needs reliable and efficient, easily computable a posteriori error estimators. A recent approach for their definition is based on the reconstruction of locally conservative fluxes in the Raviart-Thomas finite element space, yielding an a posteriori error estimator which consists only of the L^2-norm of a piecewise $H(\mathrm{div})$-vector.

Our aim is to propose a unified framework for several finite element approximations (conforming, nonconforming and discontinuous Galerkin). In this paper, we focus on the transport equation but the method can be extended to other operators.

The idea of using $H(\mathrm{div})$-reconstruction was initially proposed in [4] and has since been developed in several papers. As regards the diffusion-convection-reaction problem, the works [3] for the dG method and [5] for the mixed finite element method yield a unified approach, in which the fluxes are constructed on a dual mesh formed by dual volumes around each vertex of the primal mesh.

The main advantage of our approach is to use, contrarily to the previous references, only the primal mesh for the flux reconstruction, which presents certain facilities from a computational point of view. For this purpose, the construction

R. Becker (✉) · D. Capatina · R. Luce
EPI Concha & LMA CNRS UMR 5142 – INRIA Bordeaux Sud-Ouest & Université de Pau, IPRA, BP 1155, 64013, Pau, France
e-mail: roland.becker@univ-pau.fr; daniela.capatina@univ-pau.fr; robert.luce@univ-pau.fr

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_2,

of the $H(\mathrm{div})$-vector is inspired by the hypercircle method [1] and is achieved on patches, which may overlap and which depend on the type of the employed finite elements. The definition of the patches is related to the support of the basis functions.

In this paper, we explain the main ideas of the method and show first numerical results. In particular, we describe the construction of the a posteriori estimator for three discretizations on triangular meshes: discontinuous Galerkin, nonconforming and continuous elements. The two latter methods necessitate stabilization. For the sake of brievety, we discuss here only SUPG stabilization in the conforming case.

2 Unified Framework: The Main Ideas

We consider the model problem in a polygonal domain Ω of $\mathbb{R}^2$:

$$\begin{aligned} \boldsymbol{\beta} \cdot \nabla u &= f \text{ in } \Omega \\ u &= g \text{ on } \partial\Omega^- \end{aligned} \tag{1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^2$, $f \in L^2(\Omega)$, $g \in L^2(\partial\Omega^-)$ are given data and where the inflow boundary is defined by $\partial\Omega^- = \{x \in \partial\Omega; \boldsymbol{\beta} \cdot \mathbf{n}(x) < 0\}$. We also put $\partial\Omega^+ = \partial\Omega \setminus \partial\Omega^-$.

The weak formulation of (1) consists in finding $u \in V^g$ such that

$$a(u, v) = \int_\Omega f v \, dx, \quad \forall v \in V^0$$

where

$$a(u, v) = \int_\Omega \boldsymbol{\beta} \cdot \nabla u \, v \, dx,$$

$$V = \left\{ v \in L^2(\Omega); \boldsymbol{\beta} \cdot \nabla v \in L^2(\Omega) \right\}, \quad V^g = \left\{ v \in V;\, v_{|\partial\Omega^-} = g \right\}.$$

Let $(\mathscr{K}_h)_{h>0}$ be a regular family of triangulations of Ω consisting of triangles and let $\mathscr{S}_h^{int}$ the set of internal edges. We denote by $\mathbf{n}$ the outward unit normal to $\partial\Omega$. On any $S \in \mathscr{S}_h^{int}$ such that $\{S\} = \partial K_1 \cap \partial K_2$, we define a unit normal $\mathbf{n}_S$ and for a function ψ with $\psi_{|K_i} \in \mathscr{C}(K_i)$ we set: $\psi^{in}(\mathbf{x}) = \lim_{\varepsilon \to 0} \psi(\mathbf{x} - \varepsilon \mathbf{n}_S)$, $\psi^{ex}(\mathbf{x}) = \lim_{\varepsilon \to 0} \psi(\mathbf{x} + \varepsilon \mathbf{n}_S)$ and the jump $\left[\psi\right] = \psi^{in} - \psi^{ex}$. For $x \in \mathbb{R}$, let $x^- = \min\{0, x\}$ and $x^+ = x - x^-$. It is useful to recall the Raviart-Thomas space $RT_k = P_k^2 + \mathbf{x} P_k$, $k \in \mathbb{N}$.

We solve Eq. (1) by any of the following finite element methods: discontinuous Galerkin, nonconforming or conforming, leading to a discrete weak formulation

$$u_h \in V_h, \qquad a_h(u_h, v_h) = l_h(v_h) \quad \forall v_h \in V_h$$

where the finite dimensional space V_h and the bilinear and linear forms $a_h(\cdot,\cdot)$, $l_h(\cdot)$ will be specified later; the boundary conditions are treated by Nitsche's method.

The goal of the method is to define an approximation $\boldsymbol{\sigma}$ of $\boldsymbol{\beta}u$ satisfying

$$\boldsymbol{\sigma} \in H(\mathrm{div}, \Omega), \tag{2}$$

$$\mathrm{div}\boldsymbol{\sigma} = f_h \quad \text{in } \Omega, \tag{3}$$

$$\boldsymbol{\sigma}\cdot\mathbf{n} = \boldsymbol{\beta}\cdot\mathbf{n}g_h \quad \text{on } \partial\Omega^-, \tag{4}$$

$$\boldsymbol{\sigma}\cdot\mathbf{n} = \boldsymbol{\beta}\cdot\mathbf{n}u_h \quad \text{on } \partial\Omega^+, \tag{5}$$

where f_h and g_h are appropriate projections of the data f and g.

Actually, we do not compute $\boldsymbol{\sigma}$ but $\boldsymbol{\tau} = \boldsymbol{\sigma} - \boldsymbol{\beta}\tilde{u}_h$ with $\tilde{u}_h$ either equal to u_h itself or to a correction, depending on the employed method. The local and global a posteriori error estimators are then given by $\eta_K^2 = \|\boldsymbol{\tau}\|_{0,K}^2$ and $\eta^2 = \sum_{K\in\mathscr{K}_h}\eta_K^2 = \|\boldsymbol{\tau}\|_{0,\Omega}^2$.

The global vector $\boldsymbol{\tau}$ is obtained as the sum of local contributions $\boldsymbol{\tau}_\omega$ on (overlapping or non-overlapping) patches, a patch ω being defined as the support of a finite element basis function. On each ω, we built a piecewise $H(\mathrm{div})$-vector $\boldsymbol{\tau}_\omega$ satisfying $(\boldsymbol{\tau}_\omega)_{|K} \in RT_k$ for any $K \subset \omega$. Its computation is achieved by imposing the values of $\int_K \boldsymbol{\tau}_\omega\cdot\mathbf{r}\,dx$ for any $\mathbf{r} \in P_{k-1}^2$ if $k > 0$ and of $\int_S \boldsymbol{\tau}_\omega\cdot\mathbf{n}\varphi\,ds$ on any edge $S \subset \partial\omega$, for any basis function φ such that $S \subset \mathrm{supp}\,\varphi$. In addition, we impose certain values of $\int_K \mathrm{div}\boldsymbol{\tau}_\omega\varphi\,dx$ for $K \subset \omega$ and of $\int_S [\boldsymbol{\tau}_\omega\cdot\mathbf{n}_\mathbf{s}]\,\varphi\,ds$ on any edge S internal to the patch.

Note that there are more equations than degrees of freedom to be determined. For each finite element method, the corresponding values are computed such that the above linear system is compatible and moreover, the relations (2)–(5) hold.

As regards the error analysis, we focus here on upper bounds with respect to the following weak norm on $V + V_h$:

$$|||u||| = \sup_{v\in W} \frac{-\int_\Omega u\boldsymbol{\beta}\cdot\nabla v dx + \int_{\partial\Omega^+}\boldsymbol{\beta}\cdot\mathbf{n}uv\,ds}{\|\nabla v\|_{0,\Omega}} \tag{6}$$

where the Hilbert space W is defined by $W = \left\{v \in H^1(\Omega); v_{|\partial\Omega^-} = 0\right\} \subset V^0$.

For the discontinuous Galerkin and the nonconforming methods, we obtain by integration by parts, for any $v \in W$ and $v_h \in V_h$, that:

$$\begin{aligned} -\int_\Omega (u-u_h)\boldsymbol{\beta}\cdot\nabla v dx + \int_{\partial\Omega^+}\boldsymbol{\beta}\cdot\mathbf{n}(u-u_h)v\,ds \\ = a(u,v) - a_h(u_h,v) \\ = -\int_{\mathscr{K}_h}\boldsymbol{\tau}\cdot\nabla(v-v_h)\,dx + \int_\Omega (f-f_h)(v-v_h)\,dx. \end{aligned} \tag{7}$$

As regards the conforming method, it is well-known that it necessitates an additional stabilization term, which we denote by $s_h(\cdot,\cdot)$. The norm of the error is then modified correspondingly. In the case of the SUPG method, we can show:

$$-\int_\Omega (u-u_h)\boldsymbol{\beta}\cdot\nabla v\,dx + \int_{\partial\Omega^+} \boldsymbol{\beta}\cdot\mathbf{n}(u-u_h)v\,ds + s_h(u-u_h,v) = a(u,v) - a_h(u_h,v) + s_h(u,v) \tag{8}$$

which will be further discussed in Sect. 3.3.

Finally, for any of the three methods one deduces the upper error bound:

$$|||u-u_h||| \leq \quad \eta + ch\|f - f_h\|_{0,\Omega}. \tag{9}$$

3 Computation of Local Vectors on Patches

In what follows, we describe how to compute the vectors on a patch for the discontinuous Galerkin, the nonconforming and the conforming methods. For simplicity of presentation, we consider here $k = 1$ although it is possible to extend the methods to higher polynomial degrees. Nevertheless, the extension in the dG case being trivial, we allow for arbitrary k in this case.

3.1 Discontinuous Galerkin Method

The discrete formulation is obtained by taking V_h the space of (fully discontinuous) piecewise polynomials of degree k on any cell and

$$a_h(u_h,v_h) = \int_{\mathscr{K}_h} \boldsymbol{\beta}\cdot\nabla u_h v_h\,dx + \int_{\mathscr{F}_h^{int}} F(v_h,\mathbf{n}_S,-\boldsymbol{\beta})[u_h]\,ds + \int_{\partial\Omega^-} |\boldsymbol{\beta}\cdot\mathbf{n}|u_h v_h ds,$$

$$l_h(v_h) = \int_{\mathscr{K}_h} f v_h\,dx + \int_{\partial\Omega^-} |\boldsymbol{\beta}\cdot\mathbf{n}| g v_h\,ds,$$

where the upwind numerical flux on an internal edge S is given by:

$$F(v_h,\mathbf{n}_S,\boldsymbol{\beta}) = (\boldsymbol{\beta}\cdot\mathbf{n}_S)^+ v_h^{in} + (\boldsymbol{\beta}\cdot\mathbf{n}_S)^- v_h^{ex}.$$

For any cell K, we take $\omega_K = K$ and we define $\boldsymbol{\tau}_{\omega_K} \in RT_k$ by imposing

$$\int_K \boldsymbol{\tau}_{\omega_K} \cdot \mathbf{r}\, dx = 0, \quad \forall \mathbf{r} \in P_{k-1}^2 \quad \text{if } k > 0 \tag{10}$$

$$\int_S \boldsymbol{\tau}_{\omega_K} \cdot \mathbf{n} \varphi\, ds = -\int_S (\boldsymbol{\beta} \cdot \mathbf{n})^- [u_h] \varphi\, ds, \quad \forall S \subset \partial K, \quad \forall \varphi \in P_k. \tag{11}$$

On a boundary edge S, we put $[u_h] = 0$ if $S \subset \partial\Omega^+$ and $[u_h] = u_h - g_h$ if $S \subset \partial\Omega^-$, with g_h the piecewise P_k L^2-orthogonal projection of g. The conditions (10) and (11) correspond to the degrees of freedom of RT_k and thus completely determine $\boldsymbol{\tau}_{\omega_K}$.

By taking as test-function in the weak formulation $\varphi \chi_K$ with $\varphi \in P_k$ and χ_K the characteristics function of the cell K, it follows thanks to the previous relations that

$$\mathrm{div} \boldsymbol{\tau}_{\omega_K} = f_h - \boldsymbol{\beta} \cdot \nabla u_h \quad \text{and} \quad \left[\boldsymbol{\tau}_{\omega_K}\right] \cdot \mathbf{n}_S = -\boldsymbol{\beta} [u_h] \cdot \mathbf{n}_S,$$

for all $K \in \mathscr{K}_h$ and $S \subset \partial K$, with f_h the L^2-orthogonal projection of f on piecewise P_k elements. We finally put $\boldsymbol{\tau} = \sum_{K \in \mathscr{K}_h} \boldsymbol{\tau}_{\omega_K}$ and $\boldsymbol{\sigma} = \boldsymbol{\tau} + \boldsymbol{\beta} u_h$.

Note that (7) is obtained by writing $a_h(\cdot, \cdot)$ in the following equivalent form:

$$a_h(u_h, v_h) = -\int_{\mathscr{K}_h} u_h \boldsymbol{\beta} \cdot \nabla v_h\, dx + \int_{\mathscr{S}_h^{int}} F(u_h, \mathbf{n}_S, \boldsymbol{\beta})[v_h]\, ds + \int_{\partial\Omega^+} \boldsymbol{\beta} \cdot \mathbf{n} u_h v_h ds.$$

3.2 Nonconforming Method

We now consider V_h the space of piecewise linear functions, continuous at the midpoints of the internal edges and we take the same forms as in the dG case.

To any edge S, we associate a patch ω_S consisting of the support of the basis function φ_S associated to S. In the case of an internal edge $\{S\} = \partial K_1 \cap \partial K_2$, we have $\omega_S = K_1 \cup K_2$ and we build $\boldsymbol{\tau}_{\omega_S}$ with $(\boldsymbol{\tau}_{\omega_S})_{|K_i} \in RT_1$ for $1 \le i \le 2$ by defining the corresponding 16 degrees of freedom as follows. We impose for $i = 1, 2$:

$$\boldsymbol{\tau}_{\omega_S} \cdot \mathbf{n} = 0 \quad \text{on } \partial\omega_S \tag{12}$$

$$\int_{K_i} \boldsymbol{\tau}_{\omega_S}\, dx = \mathbf{0}, \tag{13}$$

$$\int_{K_i} \mathrm{div} \boldsymbol{\tau}_{\omega_S} \varphi_S\, dx = \int_{K_i} (f - \boldsymbol{\beta} \cdot \nabla u_h) \varphi_S\, dx + \int_{\partial K_i \setminus S} (\boldsymbol{\beta} \cdot \mathbf{n})^- [u_h] \varphi_S\, ds \tag{14}$$

$$\int_{K_i} \mathrm{div} \boldsymbol{\tau}_{\omega_S} \varphi_2^{(i)}\, dx = -\int_S (\boldsymbol{\beta} \cdot \mathbf{n})^- [u_h] \varphi_2^{(i)}\, ds. \tag{15}$$

Here above, $\varphi_2^{(i)}$ and $\varphi_3^{(i)}$ denote the two other basis functions on K_i, see Fig. 1a.

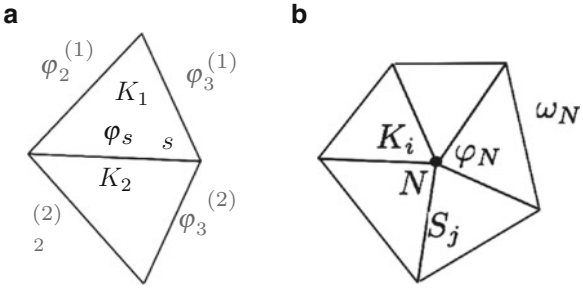


Fig. 1 Patch and basis functions for the nonconforming and conforming methods. (**a**) Nonconforming. (**b**) Conforming

Remark 1. Note that thanks to the relation $\varphi_S + \varphi_2^{(i)} + \varphi_3^{(i)} = 1$ on K_i for $1 \le i \le 2$, one can impose either $\int_{K_i} \mathrm{div}(\boldsymbol{\tau}_{\omega_S})\varphi_2^{(i)}\,dx$ or $\int_{K_i} \mathrm{div}(\boldsymbol{\tau}_{\omega_S})\varphi_3^{(i)}\,dx$ in (15).

By taking as test-function in the discrete formulation φ_S and by using that $(\varphi_S)|_S = 1$, it follows from (12), (14) and (15) that

$$\int_S [\boldsymbol{\tau}_{\omega_S} \cdot \mathbf{n}_S] ds = -\int_S \boldsymbol{\beta} \cdot \mathbf{n}_S [u_h] ds$$

$$\int_S [\boldsymbol{\tau}_{\omega_S} \cdot \mathbf{n}_S] \varphi_2^{(i)}\, ds = -\int_S \boldsymbol{\beta} \cdot \mathbf{n}_S [u_h] \varphi_2^{(i)}\, ds \quad i = 1, 2.$$

Since $\varphi_2^{(1)}{}_{|S} = \varphi_2^{(2)}{}_{|S}$ and $\left\{1, \varphi_2^{(1)}{}_{|S}\right\}$ is a basis of P_1 on S, we deduce from the two previous equalities that $[\boldsymbol{\tau}_{\omega_S} \cdot \mathbf{n}_S] = -\boldsymbol{\beta} \cdot \mathbf{n}_S [u_h]$ on S. Since $\boldsymbol{\tau} = \sum_{S \in \mathscr{S}_h} \boldsymbol{\tau}_{\omega_S}$ satisfies $[\boldsymbol{\tau} \cdot \mathbf{n}_S] = [\boldsymbol{\tau}_{\omega_S} \cdot \mathbf{n}_S]$ on any S, it follows that $\boldsymbol{\sigma} = \boldsymbol{\tau} + \boldsymbol{\beta} u_h$ belongs to $H(\mathrm{div}, \Omega)$.

3.3 Conforming Method

V_h is now the space of piecewise linear and continuous functions. For any node N we define the patch $\omega_N = \bigcup_{1 \le i \le i_N} K_i$ where $\{K_i\}_{1 \le i \le i_N}$ is the set of cells containing N as node. We consider the discrete formulation with SUPG stabilization (cf. [2]), which yields:

$$a_h(u_h, v_h) = \int_\Omega \boldsymbol{\beta} \cdot \nabla u_h v_h\, dx + \int_{\partial\Omega^-} |\boldsymbol{\beta} \cdot \mathbf{n}| u_h v_h\, ds + s_h(u_h, v_h),$$

$$s_h(u_h, v_h) = \sum_{K \in \mathscr{K}_h} \delta_K \int_K \boldsymbol{\beta} \cdot \nabla u_h \boldsymbol{\beta} \cdot \nabla v_h\, dx,$$

$$l_h(v_h) = \int_\Omega f v_h\, dx + \sum_{K \in \mathscr{K}_h} \delta_K \int_K f\, \boldsymbol{\beta} \cdot \nabla v_h + \int_{\partial\Omega^-} |\boldsymbol{\beta} \cdot \mathbf{n}| g v_h\, ds$$

and δ_K a stabilization parameter. In the case of an internal node N, we define the local vector $\boldsymbol{\tau}_{\omega_N}$ such that $(\boldsymbol{\tau}_{\omega_N})_{|K_i} \in RT_0$ for $1 \leq i \leq i_N$ and the corresponding $3\,i_N$ degrees of freedom are given by:

$$\boldsymbol{\tau}_{\omega_N} \cdot \mathbf{n} = 0 \quad \text{on } \partial\omega_N, \tag{16}$$

$$\int_{K_i} \mathrm{div}\boldsymbol{\tau}_{\omega_N} \varphi_N \, dx = \frac{1}{3} \int_{K_i} (f - \boldsymbol{\beta} \cdot \nabla u_h) \varphi_N \, dx \quad 1 \leq i \leq i_N, \tag{17}$$

$$\int_{S_i} [\boldsymbol{\tau}_{\omega_i} \cdot \mathbf{n}_{S_i}] \varphi_N \, ds = \frac{1}{2} \int_{S_i} \boldsymbol{\beta} \cdot \mathbf{n}_{S_i} [\delta_K (\boldsymbol{\beta} \cdot \nabla u_h - f_h)] \varphi_N \, ds \quad 1 \leq i \leq i_N \tag{18}$$

where S_i denotes an interior edge of the patch which has N as a node. Note that the above linear system (16)–(18) is compatible but does not have a unique solution. Indeed, let us denote by $\mathscr{A}_N$ the linear operator of the previous system. Then one can see that $\mathrm{Ker}\mathscr{A}_N$ is of dimension 1 and contains the vectors of $H(\mathrm{div}, \omega_N)$ which are divergence free, piecewise RT_0 and of zero normal trace on $\partial\omega_N$. One can also show that the kernel of $\mathscr{A}_N$ is characterized by:

$$\boldsymbol{\theta} \in \mathrm{Ker}\mathscr{A}_N \iff \int_\Omega \boldsymbol{\theta} \cdot \nabla v dx = 0, \quad \forall v \in H^1(\Omega).$$

Let us next consider the orthogonal decomposition $\boldsymbol{\tau}_{\omega_N} = \boldsymbol{\tau}_{\omega_N}^{\perp} + \boldsymbol{\tau}_{\omega_N}^{Ker}$. According to (7), it now follows that only $\boldsymbol{\tau}_{\omega_N}^{\perp}$ contributes to the error estimator η. Therefore, in order to determine it, we add to the system (16)–(18) the condition $\boldsymbol{\tau}_{\omega_N} \perp \mathrm{Ker}\mathscr{A}$.

We set $\boldsymbol{\tau} = \sum_{N \in \mathscr{N}_h} \boldsymbol{\tau}_{\omega_N}$. The numerical scheme leads to a flux $\boldsymbol{\sigma} = \boldsymbol{\tau} + \boldsymbol{\beta}\tilde{u}_h$, where $\tilde{u}_h$ represents a correction of u_h defined on each cell by $\tilde{u}_h = u_h - \delta_K(\boldsymbol{\beta} \cdot \nabla u_h - f_h)$, with f_h the piecewise constant L^2-orthogonal projection of f.

As regards the a posteriori error analysis, we obtain by imposing appropriate values of $\boldsymbol{\tau} \cdot \mathbf{n}$ on the boundary $\partial\Omega$ that

$$a(u,v) - a_h(u_h, v) + s_h(u,v) = -\int_{\mathscr{K}_h} \boldsymbol{\tau} \cdot \nabla(v - v_h) \, dx + \int_\Omega (f - f_h)(v - v_h) \, dx + \sum_{K \in \mathscr{K}_h} \delta_K \int_K (f - f_h) \boldsymbol{\beta} \cdot \nabla(v - v_h) \, dx,$$

which implies the desired estimate (9).

4 Numerical Results

We show next our first numerical results, obtained with our C++ library Concha. We consider $\Omega =]-1, 1[^2$ with data such that $u(\mathbf{x}) = \mathrm{e}^{-\frac{\|\mathbf{x}-\mathbf{x}_0\|}{\delta}}$ is an exact solution, where $\mathbf{x}_0 = (0.5, 0.5)$ and $\delta = 0.03$. We employ the dG method with $k = 0$ and

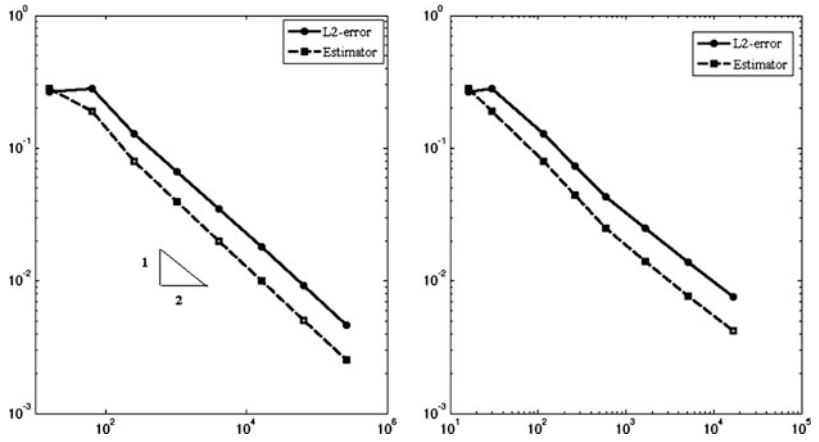


Fig. 2 L^2-error versus estimator for dG approximation ($k = 0$) in *log-log* scale (Uniform mesh (*left*), Adaptive mesh (*right*))

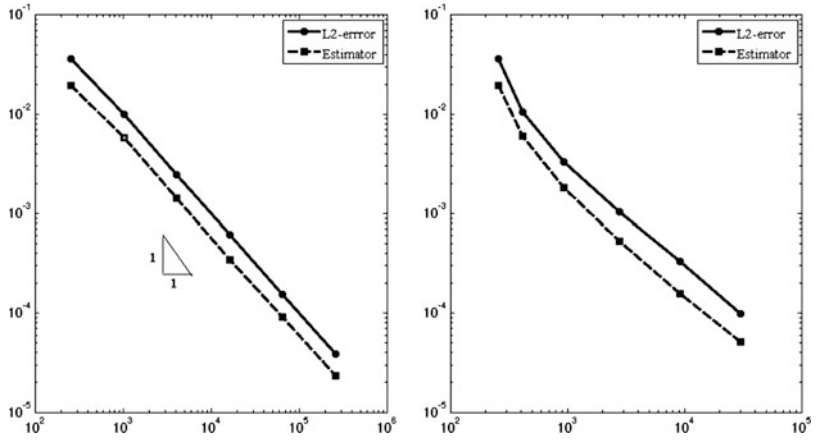


Fig. 3 L^2-error versus estimator for dG approximation ($k = 1$) in *log-log* scale (Uniform mesh (*left*), Adaptive mesh (*right*))

then $k = 1$. We represent in Figs. 2 and 3 the error and the estimator with respect to the number of cells, for uniform and adaptive mesh refinements. We have used the $L^2(\Omega)$-norm of the error since the weaker norm $|||\cdot|||$ is not easily computable. Besides the obvious gain, one can also see that both the error and the estimator converge with the same convergence rate $O(N^{-\frac{k+1}{2}})$. For $k = 1$, we show in Fig. 4 a sequence of adapted meshes. As expected, the refinement takes place near $\mathbf{x}_0$.

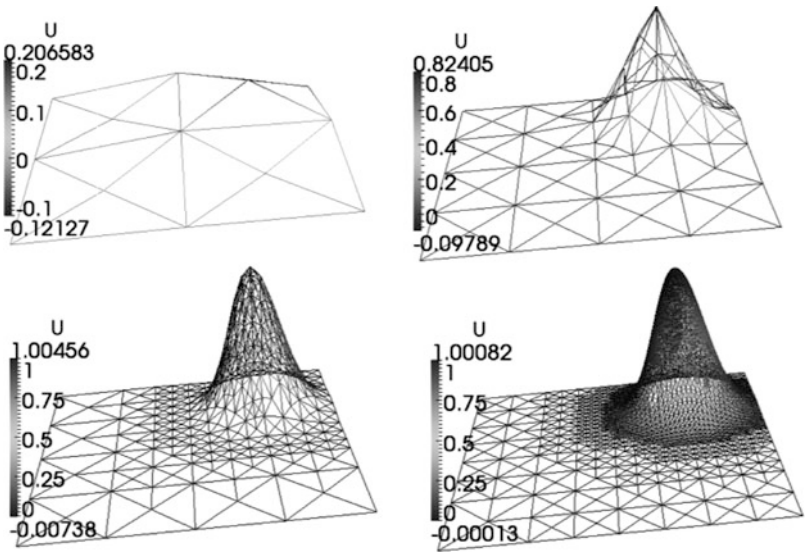


Fig. 4 Sequence of adapted meshes for dG approximation ($k = 1$)

References

1. Braess, D., Hoppe, R. H. W., Schoberl, J.: A posteriori estimators for obstacle problems by the hypercircle method. Comput. Vis. Sci. **11**, no. 4–6, 351–362 (2008)
2. Brooks, A., Hughes, T.: Streamline upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations. Comput. Methods Appl. Mech. Engrg. **31**, 199–259 (1982)
3. Ern, A., Stephansen, A. F., Vohralik M.: Guaranteed and robust discontinuous Galerkin a posteriori error estimates for convection–diffusion–reaction problems. J. Comput. Appl. Math. **234**, no. 1 (2010)
4. Luce, R., Wohlmuth, B.: A local a posteriori error estimator based on equilibrated fluxes. SIAM J. Numer. Anal. **42**, no. 4, 1394–1414 (2004)
5. Vohralik, M.: Unified primal formulation-based a priori and a posteriori error analysis of mixed finite element methods. Math. Comp. **79**, no. 272, 2001–2032 (2010)

A Posteriori Error Estimation by a Q_1/Q_2 Hierarchical Basis

M. Braack and N. Taschenberger

Abstract This work presents an a posteriori error estimation technique for Q_1 finite elements on quadrilateral triangulations by residual evaluations with respect to biquadratic test functions. The localization is performed in terms of nodal error indicators instead of cell contributions. The reliability and efficiency of the estimator is shown. Further, we discuss a simplified estimator which is even more attractive from the computational point of view.

1 Introduction

In order to design efficient numerical methods for the numerical solution of partial differential equations by using adaptive meshes the adaptation criteria is very important. Therefore, a posteriori error estimation and local error indicators play a crucial role.

The first a posteriori error estimation techniques for finite element discretizations have been developed by Babuška and Rheinboldt [3]. Since then, several alternative approaches have been proposed and analyzed, e.g., residual based indicators [1] and hierarchical estimators [4,9]. Moreover, the books of Ainsworth and Oden [2] and of Babuška and Strouboulis [7] include overviews of different techniques. The work of Verfürth [8] was an important step because it was not only shown that the proposed estimator is reliable but also efficient. That means that the estimator can be bounded by the discretitzation error multiplied by a mesh size independent constant.

M. Braack (✉) · N. Taschenberger
Mathematisches Seminar, Christian-Albrechts-Universität zu Kiel, Ludewig-Meyn-Str. 4, D-24098, Kiel, Germany
e-mail: braack@math.uni-kiel.de; taschenberger@math.uni-kiel.de

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_3,

We propose in this work a posteriori error estimators for bilinear finite elements based on the evaluation of residuals with respect to biquadratic test functions. From the practical point of view, the estimator has the advantage that the computation of jump terms is not necessary. This is in particular advantageous on quadrilateral meshes with hanging nodes. We show locally equivalence of the estimator and the "standard estimator" of [8]. A numerical example illustrates the behaviour in practice and shows the reliability and efficiency.

2 Notations

Let $\Omega \subset I\!R^2$ be a two-dimensional polygonal domain. For any open subset $\omega \subset \Omega$ let $L^2(\omega)$ be the Lebesgue space of square-integrable functions over ω, and $H^k(\omega)$ the Sobolev space with weak derivatives up to order $k \in I\!N$. The corresponding norms are denoted by $\|\cdot\|_\omega$ and $\|\cdot\|_{k;\omega}$, respectively. The L^2-scalar product and norm is denoted by $(\cdot,\cdot)_\omega$ and $\|\cdot\|_\omega$, respectively. In the case $\omega = \Omega$, we simply use $\|\cdot\|$, $\|\cdot\|_k$ and $(\cdot,\cdot)$. Furthermore, the Hilbert space of H^1 functions with vanishing traces on the boundary is denoted by $V := H_0^1(\Omega) = \{\phi \in H^1(\Omega) : \phi = 0$ a.e. on $\partial\Omega\}$.

Let $\{\mathscr{T}_h\}_{h>0}$ be a shape regular family of triangulations of Ω consisting of quadrilaterals. For given h and $T \in \mathscr{T}_h$, h_T and ρ_T denote the diameter and the inner radius of T, respectively. The set of internal edges of $\mathscr{T}_h$ will be denoted by $\mathscr{E}_h$, i.e. for each edge $e \in \mathscr{E}_h$ the intersection $e \cap \partial\Omega$ contains at most two boundary points. The shape regularity implies that the diameters $h_T, h_{T'}$ of two neighbouring cells $T, T' \in \mathscr{T}_h$ and the length h_e of a neighbouring edge e scale similar up to a h-independent constant $c \geq 1$:

$$0 \;<\; \max\{h_T, h_{T'}, h_e\} \leq c \min\{h_T, h_{T'}, h_e\}\,.$$

We use the space of polynomials up to total degree r, denoted by Q_r. The finite element space is

$$V_h^{(r)} := \{\phi \in V : \phi|_T \in Q_r\,, \forall T \subset \mathscr{T}_h\}\,.$$

The space of bilinear elements is simply denoted by $V_h := V_h^{(1)}$. By $\mathscr{N}_h$ we denote all inner nodes of $\mathscr{T}_h$, i.e. those nodes which are not located on the (Dirichlet) boundary $\partial\Omega$. We set $n_h := |\mathscr{N}_h| = \dim V_h$. The notation $a \prec b$ stands for $a \leq cb$ with a h-independent constant $c > 0$.

Furthermore, we use the L^2-projection $\pi_h : L^2(\Omega) \to \{\phi \in L^2(\Omega) : \phi|_T \in Q_0\,, \forall T \subset \mathscr{T}_h\}$, and $[v]_e$ for the jump of a cell-wise polynomial v across an edge e of $\mathscr{T}_h$. For edges e on the (Dirichlet) boundary, we make the convention $[v]_e := 0$.

3 Standard a Posteriori Energy Estimate for the Poisson Problem

We consider the Poisson problem with homogeneous Dirtichlet conditions. All results carry over to mixed Dirichlet-Neumann conditions with the usual modifications. We seek for given right hand side $f \in L^2(\Omega)$ the function $u \in V$ such that

$$(\nabla u, \nabla \phi) = (f, \phi) \qquad \forall \phi \in V\,.$$

Due to the Theorem of Riesz, there is always a unique solution. The corresponding finite element formulation reads

$$u_h \in V_h : \ (\nabla u_h, \nabla \phi) = (f, \phi) \quad \forall \phi \in V_h\,. \tag{1}$$

The standard a posteriori energy error estimate of [8] (on triangular meshes and P_1-elements) consists of cell-wise error contributions

$$\eta_{SEE} := \left(\sum_{T \in \mathcal{T}_h} \eta_T^2 \right)^{1/2}, \tag{2}$$

$$\eta_T := \left(h_T^2 \|\pi_h f + \Delta u_h\|_T^2 + h_T \|[\partial_n u_h]\|_{\partial T}^2 \right)^{1/2}.$$

According to [8], there are constants c_1, c_2 depending only on the polynomial degree r and the shape regularity of the family $\{\mathcal{T}_h\}_{h>0}$ such that

$$\|\nabla(u - u_h)\|^2 \prec \eta_{SEE}^2 + \sum_{T \in \mathcal{T}_h} h_T^2 \|f - \pi_h f\|_T^2\,, \tag{3}$$

$$\eta_T^2 \prec \|\nabla(u - u_h)\|_{\omega_T}^2 + h_T^2 \|f - \pi_h f\|_{\omega_T}^2 \qquad \forall T \in \mathcal{T}_h\,. \tag{4}$$

Here, we use the notation ω_T for a patch of cells $T' \in \mathcal{T}_h$ having at least one edge in common with T. If the discrete solution u_h is known, the local error indicators η_T can be computed and used for local mesh refinement. According to [5], the two contributions of η_T (the cell residual and the jump of the derivatives) are not equilibrated. For instance, for low-order finite elements (linear or bilinear), the jump term dominates the cell residual term; see [9, 10] for details.

4 An Error Estimator Based on Higher-Order Residuals

We present an alternative approach with error indicators related to the degrees of freedom and not to the cells. For Q_1-elements the degrees of freedom are directly related to the inner nodes of the mesh. Instead of evaluating jump terms,

the estimator just consists in computing residuals to higher-order test functions. The construction takes advantage of the hierarchical basis of $V_h^{(r+1)}$ and is related to the approach on triangular meshes in [9]. We refer to Bank [4] for further benefits of hierarchical bases for finite elements.

Let $\mathscr{B}_h^{(1)} := \{\psi_1, \ldots, \psi_{n_h}\}$ denote the Lagrangian nodal basis of V_h. Each ψ_i is a standard 'hat-function' corresponding to a node $N_i \in \mathscr{N}_h$. In order to build a biquadratic basis $\mathscr{B}_h^{(2)}$ of $V_h^{(2)}$, we consider the mesh $\mathscr{T}_{h/2}$ which arises by one global refinement of $\mathscr{T}_h$. We divide the nodes of $\mathscr{N}_{h/2}$ in three types,

$$\mathscr{N}_{h/2} = \mathscr{N}_h \cup \mathscr{N}_{h/2}^E \cup \mathscr{N}_{h/2}^C ,$$

where $\mathscr{N}_{h/2}^E$ consists of nodes located on inner edges of $\mathscr{T}_h$, and $\mathscr{N}_{h/2}^C$ consists of nodes located on cell centers of $\mathscr{T}_h$. Now, the hierarchical basis $\mathscr{B}_h^{(2)}$ consists of the hat-functions $\psi \in \mathscr{B}_h^{(1)}$, edge bubbles ψ_e, and cell bubbles ψ_T,

$$\mathscr{B}_h^{(2)} := \mathscr{B}_h^{(1)} \cup \{\psi_e : e \in \mathscr{E}_h\} \cup \{\psi_T : T \in \mathscr{T}_h\} .$$

This leads to a canonical bijective mapping between the indices and this quadratic basis, $\{1, \ldots, n_{h/2}\} \to \mathscr{B}_h^{(2)}$, $k \mapsto \psi_k^{(2)}$, which corresponds to the numbering of the nodes of $\mathscr{N}_{h/2}$. This means, if $k \in \mathscr{N}_{h/2}^C$, then $\psi_k^{(2)}$ is the cell bubble function centered at N_k; if $k \in \mathscr{N}_{h/2}^E$, then $\psi_k^{(2)}$ is the edge bubble function centered at the node N_k. For each of these indices $1 \le i \le n_{h/2}$, we define the residual Φ_i corresponding to this quadratic basis and their sum will be the **H**igher-order a posteriori **E**rror **E**stimator:

$$\Phi_i := (\pi_h f, \psi_i^{(2)}) - (\nabla u_h, \nabla \psi_i^{(2)}) , \tag{5}$$

$$\eta_{HEE} := \left(\sum_{i=1}^{n_{h/2}} \Phi_i^2 \right)^{1/2} . \tag{6}$$

4.1 Local Equivalence of the Error Estimators

We will show that the cell error indicators (2) and the nodal based error indicators $|\Phi_i|$ are "locally equivalent" in the sense of Dörfler [6]. For each cell $T \in \mathscr{T}_h$ the set $I(T) \subset \{1, \ldots, n_{h/2}\}$ contains the indices of the corresponding "cell bubble" function and "edge bubble" functions; $|I(T)| \le 5$. In the opposite way, for each node $N_i \in \mathscr{N}_{h/2}$, we define $T(i) \subset \mathscr{T}_h$, so that $T \in T(i)$ implies $N_i \in T$. We are now in the situation to show the local equivalence between η_{SEE} and η_{HEE}.

Proposition 1. *Let $\{\mathcal{T}_h\}_{h>0}$ be a shape-regular family of parallelogram meshes. Then the cell based indicators (2) and the nodal based indicators $|\Phi_i|$ are locally equivalent, i.e.,*

$$0 \leq |\Phi_i| \prec \sum_{T \in T(i)} \eta_T \quad \text{and} \quad 0 \leq \eta_T \prec \sum_{i \in I(T)} |\Phi_i| \,. \tag{7}$$

Proof. The assertions follow immediately from the following estimates:

(a) For $N_k \in \mathcal{N}_{h/2}^C$ let $T \in \mathcal{T}_h$ be the corresponding cell (i.e. $N_k \in T \setminus \partial T$). It holds that

$$\begin{aligned} & |\Phi_k| = \left|(\pi_h f + \Delta u_h)|_T\right| \int_T \psi_k^{(2)} dx \;=\; T^{-1/2} \|\pi_h f + \Delta u_h\|_T \int_T \psi_k^{(2)} dx \\ \Longrightarrow \quad & |\Phi_k| \prec h_T \|\pi_h f + \Delta u_h\|_T \prec |\Phi_k| \,. \end{aligned}$$

(b) For $N_j \in \mathcal{N}_{h/2}^E$ on an edge $e \in \mathcal{E}_h$ let $N_i, N_k \in \mathcal{N}_{h/2}^C$ be the nodes of the adjacent cell bubble functions. It holds that

$$\begin{aligned} h_e^{1/2} \|[\partial_n u_h]\|_e \prec |(\nabla u_h, \nabla \psi_j^{(2)})| \;&=\; |(\pi_h f + \nabla u_h, \psi_j^{(2)}) - \Phi_j| \\ &\leq \|\pi_h f + \nabla u_h\|_{T_i \cup T_k} |T_i \cup T_k|^{1/2} + |\Phi_j| \;\leq\; |\Phi_i| + |\Phi_j| + |\Phi_k| \end{aligned}$$

and $$|\Phi_j| \prec h_{T_i} \|\pi_h f + \Delta u_h\|_{T_i \cup T_k} + h_e^{1/2} \|[\partial_n u_h]\|_e \,.$$

Proposition 2. *Let $\{\mathcal{T}_h\}_{h>0}$ be a shape-regular family of parallelogram meshes. Then it holds for the estimator given by (6)*

$$\|\nabla(u - u_h)\|^2 \prec \eta_{HEE}^2 + \sum_{T \in \mathcal{T}_h} h_T^2 \|f - \pi_h f\|_T^2 \,, \tag{8}$$

$$|\Phi_i| \prec \|\nabla(u - u_h)\|_{T(i)} \qquad \forall i \in \{1, \ldots, n_{h/2}\} \,. \tag{9}$$

Proof. The reliability bound (8) is a direct consequence of the previous proposition and (3) and (4). The efficiency bound (9) cannot be derived by the previous propositions due to the appearance of fluctuations in the right hand side, $\|f - \pi_h f\|_T$ in (3). Therefore, we use the criterion in [9] (Propositions 1.14 and 1.15) which is originally formulated for triangular meshes and P_1-elements but can easily be generalized to Q_1-elements on quadrilateral meshes. To this end, we observe

that for each edge bubble function $\psi_e \in \mathscr{B}_h^{(2)}$ and all cell bubble basis functions $\psi_T \in \mathscr{B}_h^{(2)}$ it holds

$$0 \le \psi_e, \psi_T \ \le \ 1 \, .$$

Furthermore, let $e \in \mathscr{E}_h$ be an inner edge and $T_i, T_k \in \mathscr{T}_h$ the adjacent cells. Then it is easy to verify by a simple scaling argument that the following estimates for the corresponding edge bubble function $\psi_e \in \mathscr{B}_h^{(2)}$ hold:

$$c h_e \le \int_e \psi_e \, dx \, , \qquad h_e \|\nabla \psi_e\|_{T_i \cup T_k} \ \prec \ \|\psi_e\|_{T_i \cup T_k} \, ,$$

with $c > 0$. Similarly, it holds for each cell bubble basis function $\psi_T \in \mathscr{B}_h^{(2)}$:

$$h_T^2 \le \int_T \psi_T \, dx \, , \quad h_T \|\nabla \psi_T\|_T \ \prec \ \|\psi_T\|_T \, .$$

Now, the bound (9) follows by applying Propositions 1.14 and 1.15 of [9].

4.2 A Cheaper Variant on a Coarser Mesh

Although the proposed estimator (6) does not need the solution of local problems, the evaluation of residual terms for the higher-order test functions is more expensive than the residual of the lower-order test functions. For reducing the costs further, one may change the estimator by taking the higher-order residuals on a coarser mesh. For this we assume that the triangulation $\mathscr{T}_h$ is organized in a patch-wise manner, i.e., $\mathscr{T}_h$ results from a coarser locally refined mesh $\mathscr{T}_{2h}$ by one global refinement. An Coarse Error Estimator (CEE) can now be formulated as before, but with the difference that the quadratic test functions $\psi_i^{(2)}$ are chosen out of the space $V_{2h}^{(2)}$,

$$\Phi_i' := (\pi_h f, \psi_i^{(2)}) - (\nabla u_h, \nabla \psi_i^{(2)}) \, , \quad \psi_i^{(2)} \in \mathscr{B}_{2h}^{(2)} \, ,$$

$$\eta_{CEE} := \left(\sum_{i=1}^{n_h} |\Phi_i'|^2 \right)^{1/2} \, .$$

Only the nodes of $\mathscr{N}_h$, and hence much less quadratic test functions are considered. Note that $\Phi_i' = 0$ if $N_i \in \mathscr{N}_{2h}$, because $\psi_i^{(2)} \in V_h$ due to the hierarchical construction of the basis. The local indicators $|\Phi_i'|$ can be bounded from above by some $|\Phi_j|$:

Proposition 3. *It holds*

$$|\Phi_i'| \prec \sum_{j \in I_i} |\Phi_j| \prec \left(\|\nabla(u-u_h)\|^2_{T(i)} + \sum_{T \in T(i)} h_T^2 \|f - \pi_h f\|^2_T \right)^{1/2} .$$

Proof. Since $V_{2h}^{(2)} \subset V_h^{(2)}$, the quadratic test function $\psi_i^{(2)} \in \mathscr{B}_{2h}^{(2)}$ can be expressed by a linear combination of test functions in $\mathscr{B}_h^{(2)}$ with support in $T(i)$: In particular, there is a finite number of h-independent coefficients $\{\beta_{ji} : j \in I_j\}$ such that

$$|\Phi_i'| = \left| \sum_{j \in I_i} \beta_{ji} \Phi_j \right| \leq \sum_{j \in I_i} |\beta_{ji}||\Phi_j| .$$

The second upper bound follows immediately from Proposition 2 and the fact that for $j \in I_i$, $T(j) \subset T(i)$.

5 Numerical Results

We perform a numerical comparison between the estimators η_{SEE}, η_{HEE} and η_{CEE}. An important and established quality measure is the efficiency index

$$I_{\mathit{eff}}(\eta) = \frac{\eta}{\|\nabla(u-u_h)\|} .$$

We consider the slit domain $\Omega = (-1,1)^2 \setminus \{(x_1,x_2) \,|\, x_1 \in [0,1], x_2 = 0\}$. The right-hand side is equal to one, $f \equiv 1$, and the boundary conditions are homogeneous Dirichlet conditions. Due to Galerkin orthogonality it holds $\|\nabla(u-u_h)\| = \sqrt{\|\nabla u\|^2 - \|\nabla u_h\|^2}$. The value $\|\nabla u\|^2$ is approximated numerically by a Q_2 solution on a very fine mesh with approximately four millions of cells, $\|\nabla u\|^2 \approx e^2 = 0.32438$. It holds $|\|\nabla u\|^2 - e^2| \leq 10^{-4}$.

Figure 1 shows the estimated error and the obtained efficiency indices on globally refined meshes. The efficiency of the standard error estimator SEE is between 4.3 and 2.4 and is getting better on finer meshes. For HEE the efficiency is close 1.7 for all meshes, the efficiency index of CEE is always between 0.58 and 0.70, the error is underestimated on all meshes but the asymptotic behaviour is extremely good. The obtained efficiency indices on locally refined meshes are shown in Fig. 2. On locally refined meshes, all of them seem to be slightly less robust than on structured meshes. However, the cheap variant CEE delivers the best results in the sense that they are the closest to one. Moreover, the asymptotic behavior of CEE is very stable.

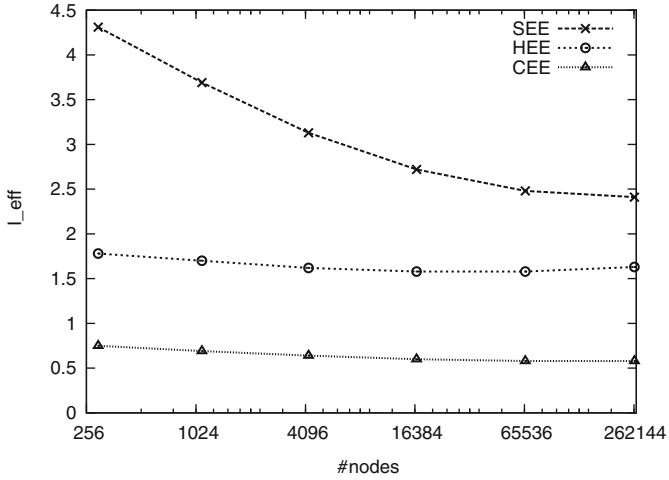


Fig. 1 Comparison of efficiency indices of the different estimators on globally refined meshes

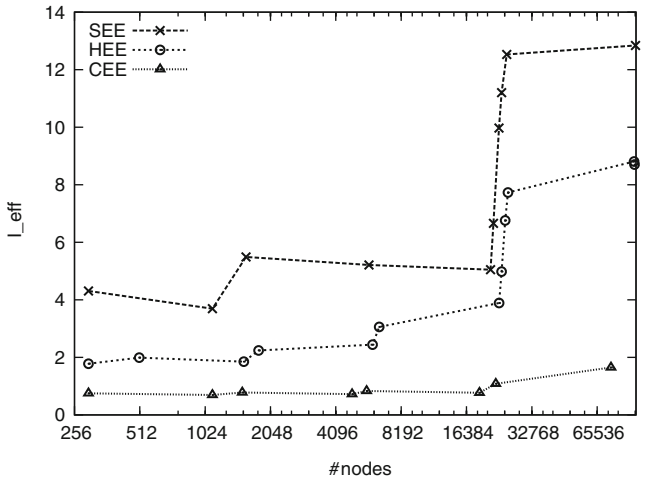


Fig. 2 Comparison of efficiency indices of the different estimators on locally refined meshes

Finally, we compare in Fig. 3 the relation between n_h (number of degrees of freedom) and $\|\nabla(u - u_h)\|$ on globally refined meshes and on locally refined meshes obtained by the different estimators. In this context, the adaptive strategy outperforms global refinement independently of the type of estimator. But even though the estimated quantities differ, the resulting local refined meshes lead to the same convergence rates for all three error estimators.

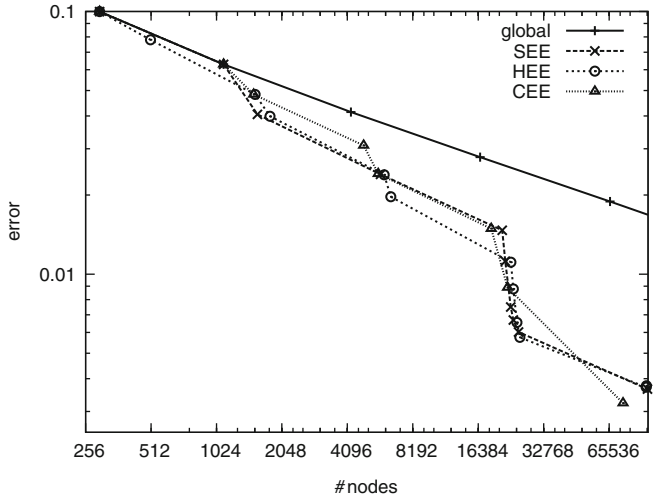


Fig. 3 Comparison of performance of the different estimators

Acknowledgements We gratefully acknowledge the partial support of this work by the DFG Priority Program SPP 1276 (MetStröm).

References

1. M. Ainsworth and J. T. Oden. A unified approach to a posteriori error estimation using element residual methods. *Numer. Math.*, 65(1):23–50, 1993.
2. M. Ainsworth and J. T. Oden. *A Posteriori Error Estimation in Finite Element Analysis*. Wiley Interscience, New York, 2000.
3. I. Babuska and W. C. Rheinboldt. Error estimates for adaptive finite element computations. *SIAM J. Numer. Anal.*, 15:736–754, 1978.
4. R.E. Bank. Hierarchical bases and the finite element method. *Acta Numer.*, 44:1–43, 1996.
5. C. Carstensen and R. Verfürth. Edge residual dominate a posteriori error estimates for low order finite element methods. *SIAM J. Numer. Anal.*, 36(5):1571–1587, 1999.
6. W. Dörfler. A convergent adaptive algorithm for Poisson's equation. *SIAM J. Numer. Anal.*, 33(3):1106–1124, 1996.
7. I. Babuška and T. Strouboulis. *The Finite Element Method and its Reliability*. Oxford Univers. Press, 2001.
8. R. Verfürth. A posteriori error estimates for nonlinear problems. finite element discretization of elliptic equations. *Math. Comput.*, 62:445–475, 1994.
9. R. Verfürth. *A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*. John Wiley/Teubner, New York-Stuttgart, 1996.
10. D. Yu. Asymptotically exact a posteriori error estimators for elements of bi-odd degree. *Chinese J. Num. Math. and Appl.*, 13:64–78, 1991.

Adaptive Finite Elements with Anisotropic Meshes

W. Huang, L. Kamenski, and J. Lang

Abstract The paper presents a numerical study for the finite element method with anisotropic meshes. We compare the accuracy of the numerical solutions on quasi-uniform, isotropic, and anisotropic meshes for a test problem which combines several difficulties of a corner singularity, a peak, a boundary layer, and a wavefront. Numerical experiment clearly shows the advantage of anisotropic mesh adaptation. The conditioning of the resulting linear equation system is addressed as well. In particular, it is shown that the conditioning with adaptive anisotropic meshes is not as bad as generally assumed.

1 Introduction

Anisotropic mesh adaptation, i.e., adaptation of the size and shape of mesh elements, has been shown to be of significant advantage for problems with distinct anisotropic features. Moreover, the ability to adjust the shape and orientation of mesh elements has proven to be useful for designing numerical schemes with particular features, e.g., satisfying the discrete maximum principle [23] or improving the conditioning of the finite element equations [7, 22].

In this paper we concentrate on obtaining anisotropic meshes for the purpose of minimizing the numerical solution error. Typically, the optimal shape and orientation of mesh elements depend on the Hessian [6, 12, 18, 27] or the first derivatives [25, 26] of the exact solution of the underlying problem. This is the first

W. Huang · L. Kamenski (✉)
Department of Mathematics, University of Kansas, Lawrence, USA
e-mail: huang@math.ku.edu; lkamenski@math.ku.edu

J. Lang
Graduate School of Computational Engineering, Center of Smart Interfaces, Department of Mathematics, Technische Universität Darmstadt, Darmstadt, Germany
e-mail: lang@mathematik.tu-darmstadt.de

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_4, © Springer-Verlag Berlin Heidelberg 2013

major difficulty since the exact solution is usually not available. One possibility to solve this difficulty is to try to recover the approximate Hessian from the numerical solution in the course of the computation. In [10, 12], mesh adaptation is based on a residual-based error estimator but still requires Hessian recovery for the solution of the dual problem. Unfortunately, Hessian recovery methods do work very well for interpolation problems but they cannot provide a convergent recovery if applied to linear finite element approximations on non-uniform meshes [3, 21], although adaptive finite element methods based on Hessian recovery still provide excellent mesh adaptation in practice [9, 10, 12, 20, 23, 28]. The fact that the convergence of adaptive algorithms employing Hessian recovery cannot be proven directly (since the recovered Hessian does not converge to the exact Hessian) explains recent interest in anisotropic mesh adaptation based on some kind of a posteriori error estimates which do not depend on the exact solution of the underlying problem [1–5, 13, 14, 19]. Moreover, as shown in [19, Sect. 5.3], using error estimates could be of advantage for problems exhibiting gradient jumps or similar discontinuities along internal interfaces because methods based on recovery of derivatives could result in unnecessarily high mesh density near discontinuities.

For our study we employ the anisotropic mesh adaptation algorithm from [19] which employs a globally defined hierarchical basis error estimate (HBEE) for obtaining the directional information. In contrast to the recovery-based algorithms, this method adapts the mesh in order to directly minimize the a posteriori error estimate and, thus, relies on the accuracy of the error estimator but does not require recovery of derivatives of the exact solution. In this sense, the algorithm is completely a posteriori.

Another major concern when using anisotropic meshes is the conditioning of the finite element equations. Generally speaking, an anisotropic mesh is expected to contain elements of large aspect ratio[1] and there exists a concern that an anisotropic mesh will lead to extremely ill-conditioned linear algebraic systems and this may weaken the accuracy improvements gained through anisotropic mesh adaptation. Fortunately, as it has been recently shown in [22], the conditioning of the stiffness matrix with anisotropic meshes is not necessarily as bad as generally assumed, especially in $2D$. In Sect. 3.2, we will see that even if the condition number of the stiffness matrix with an anisotropic mesh is larger than that with an isotropic mesh, the accuracy gained through anisotropic mesh adaptation still clearly outbalances the conditioning issues, at least for the example considered.

The outline of this paper is as follows: a brief description of the adaptation algorithm is given in Sect. 2 which is followed by the numerical experiment in Sect. 3. The concluding remarks are given in Sect. 4.

[1] In this paper the aspect ratio of a triangular element is defined as the longest edge divided by the shortest altitude. For example, an equilateral triangle has an aspect ratio of $2/\sqrt{3} \approx 1.15$.

2 Discretization and the Mesh Adaptation Algorithm

We consider a Dirichlet problem for the Poisson equation

$$\begin{cases} -\Delta u = f, & \text{in } \Omega \\ u = 0, & \text{on } \partial\Omega \end{cases} \tag{1}$$

where $\Omega \subset \mathbb{R}^2$ is a connected bounded polygonal domain.

For a given triangulation $\mathscr{T}_h$ of Ω and the associated linear finite element space $V_h \subset H_0^1(\Omega)$, the linear finite element solution $u_h \in V_h$ of (1) is defined by

$$\int_\Omega \nabla v_h \cdot \nabla u_h \, dx = \int_\Omega f v_h \, dx, \quad \forall v_h \in V_h. \tag{2}$$

The finite element space V_h and the finite element solution u_h can be written as

$$V_h = \text{span}\{\phi_1, \cdots, \phi_{n_{int}}\} \quad \text{and} \quad u_h = \sum_{j=1}^{n_{int}} u_j \phi_j, \tag{3}$$

where ϕ_j is the standard linear basis function associated with the j-th vertex and n_{int} is the number of interior vertices of the triangulation. Substituting (3) into (2) and taking $v_h = \phi_i$ for $i = 1, \ldots, n_{int}$ results in the linear system

$$Au_h = F, \tag{4}$$

where

$$A_{ij} = \int_\Omega \nabla\phi_j \cdot \nabla\phi_i \, dx \quad \text{and} \quad F_i = \int_\Omega f\phi_i \, dx.$$

Note that in order to obtain the finite element solution u_h we need to solve the linear algebraic system (4). Thus, the accuracy of u_h depends also on the conditioning of this system which in turn is affected by the choice of the mesh. As mentioned in the introduction, there is a concern that anisotropic meshes could lead to extremely ill-conditioned linear systems and this may weaken the accuracy gained with anisotropic mesh adaptation. In our numerical experiment in Sect. 3.2 we will address this issue in detail.

In order to construct $\mathscr{T}_h$ (and, thus, the corresponding V_h) we employ the M-uniform mesh approach which generates an adaptive mesh as a quasi-uniform one in the metric specified by a symmetric and strictly positive definite tensor $M = M(x)$ [17]. The algorithm starts with an initial mesh. For every mesh $\mathscr{T}_h^{(i)}$ we compute the finite element solution $u_h^{(i)}$ which is used to compute a new adaptive mesh for the next iteration step. The new mesh is generated as an M-uniform mesh with a metric tensor $M_h^{(i)}$ computed from $u_h^{(i)}$. This yields the sequence

$$\mathscr{T}_h^{(0)} \to u_h^{(0)} \to M_h^{(0)} \to \mathscr{T}_h^{(1)} \to u_h^{(1)} \to M_h^{(1)} \to \dots$$

The mesh adaptation process is repeated until the mesh is M-uniform within a given tolerance (see [19, Sect. 4.1] for more details). In our computation we use BAMG (*bidimensional anisotropic mesh generator* [16]) to construct anisotropic meshes for a given metric tensor M.

Typically, the optimal metric tensor M_h depends on the Hessian of the exact solution [11, 12, 17] which is usually unknown. In this study we follow [19] and employ the hierarchical basis a posteriori error estimate (HBEE) to obtain the directional information required for the metric tensor M_h. The brief idea is as follows (see [19] for details).

If we have an error estimate z_h such that

$$\|u - u_h\| \le C\,\|z_h\|.$$

for a given norm $\|\cdot\|$ and if it further has the property $\Pi_h z_h \equiv 0$ with Π_h being the interpolation operator associated with V_h (which is fulfilled by the HBEE), than the finite element approximation error is bounded by the interpolation error of the error estimate,

$$\|u - u_h\| \le C\,\|z_h\| = C\,\|z_h - \Pi_h z_h\|. \tag{5}$$

Hence, up to a constant, the solution error is bounded by the interpolation error of the error estimate and the mesh can be constructed to minimize the interpolation error of z_h; the metric tensor M_h does not depend on the Hessian of the exact solution.

In this study, we are concerned with the error measured in the H^1 semi-norm, which is the energy norm from (1). Therefore, instead of using the metric tensor developed in [19] for the error measured in the L^2 norm, we construct the metric tensor which minimizes the interpolation error of z_h measured in the H^1 semi-norm. In two dimensions the optimal metric tensor is given element-wise by

$$M_K = \left\| I + \frac{1}{\alpha_h}|H_K(z_h)| \right\| \cdot \det\left(I + \frac{1}{\alpha_h}|H_K(z_h)| \right)^{-\frac{1}{4}} \cdot \left[I + \frac{1}{\alpha_h}|H_K(z_h)| \right],$$

where $H_K(z_h)$ denotes the Hessian of the (quadratic) hierarchical basis error estimate z_h on element K and α_h is a regularization parameter to ensure that M_K is strictly positive definite. α_h can also be seen as an adaptation intensity control: uniform mesh has $\alpha_h = \infty$ and if $\alpha_h \to 0$ the mesh becomes more adaptive. Usually, α_h is chosen so that about half of the mesh elements are concentrated in regions where $\det(M)$ is large (see [18] for more details on the choice of M_K and α_h).

In our computations we employ the globally defined hierarchical basis error estimate since it contains more directional information of the solution than localized versions [19, Sect. 5.1]. Moreover, it has been shown that local error estimates can be inaccurate on anisotropic meshes [8]. To avoid the cost of the exact solution of the global error problem, we use only a few sweeps of the symmetric Gauss-Seidel iteration for the resulting linear system until the relative difference of the old and

the new error approximations is under a given relative tolerance. This proves to be adequate for the purpose of mesh adaptation and the computational cost is comparable to that of the Hessian recovery: in the tests, the computation of HBEE is about twice slower than Hessian recovery.

Although the validity of the classical hierarchical basis error estimate z_h for the anisotropic case is still unclear, theoretical considerations in [15, Sect. 6.4] and numerical results in [19] suggest that the hierarchical basis error estimate is a reliable source of information when a mesh is aligned with the solution.

3 Numerical Experiment

For the numerical experiment we consider a problem in [24] which combines multiple difficulties. It is a Dirichlet problem of the Poisson equation

$$\begin{cases} -\Delta u = f, & \text{in } \Omega \\ u = g, & \text{on } \partial\Omega \end{cases} \tag{6}$$

where Ω is an L-shaped domain $\Omega = (-1,1) \times (-1,1) \setminus [0,1) \times (-1,0]$. The functions f and g are chosen such that the exact solution u is given by

$$u(x,y) = r^{2/3}\sin(2\theta/3) + \tan^{-1}\left(200\left(\sqrt{x^2+(y+3/4)^2} - 3/4\right)\right) + e^{-1000\left((x+\sqrt{5}/4)^2+(y+1/4)^2\right)} + e^{-100(y+1)},$$

where r and θ are the polar coordinates. The solution has

- A singular gradient at $(0,0)$ due to a reentrant corner of the L-shaped domain Ω,
- A circular wavefront with the center in $(0,-3/4)$ and the radius of $3/4$,
- A sharp peak at $(-\sqrt{5}/4,-1/4)$,
- And a boundary layer along the line $y=-1$.

Figure 1 shows the surface and the color plot of a numerical solution.

3.1 Accuracy of the Numerical Solution

First, we compare the accuracy of the numerical solution for Delaunay (quasi-uniform), adaptive isotropic, and adaptive anisotropic meshes. Examples of mesh types are given in Fig. 2. We observe that both isotropic and anisotropic adaptive meshes (Fig. 2b, c, respectively) have high mesh density in regions with difficulties but the anisotropic mesh (Fig. 2c) is clearly much better aligned with the steep

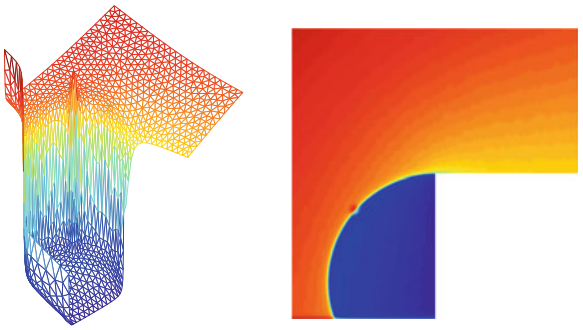


Fig. 1 Surface and color plots of the numerical solution

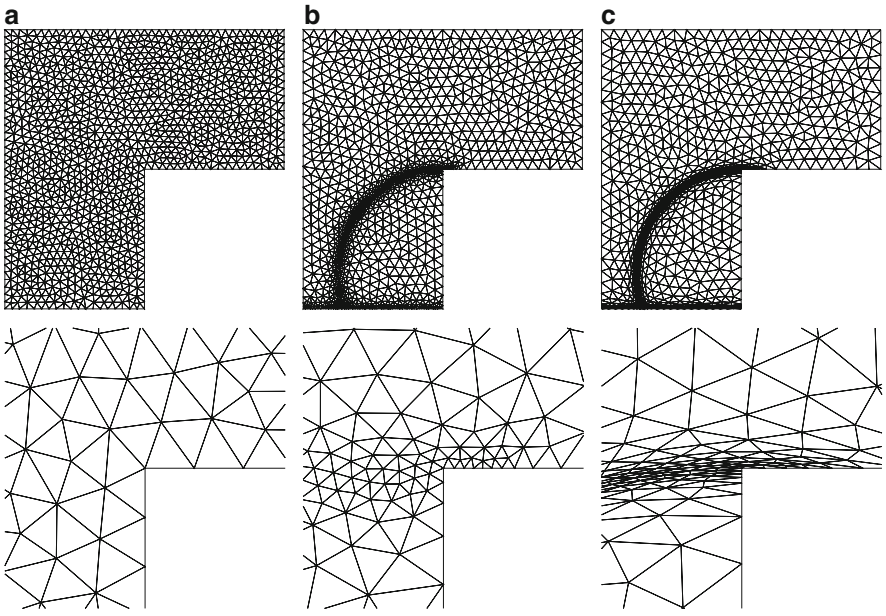


Fig. 2 Mesh examples and 6.6 times close-up views at the reentrant corner. (**a**) Delaunay: 2326 elements; max. aspect ratio 2.8. (**b**) Isotropic adaptive: 2321 elements; max. aspect ratio 3.0. (**c**) Anisotropic adaptive: 2316 elements; max. aspect ratio 24.4

boundary layer and the wavefront. This is the major difference between the isotropic and anisotropic adaptation: the isotropic adaptation can provide proper mesh density whereas the anisotropic adaptation can provide both proper mesh density *and* proper alignment of the mesh with the anisotropic features of the solution.

Figure 3 shows the error of the numerical solution measured in the energy norm $|||u - u_h|||$, which is equal to $H^1(\Omega)$ semi-norm $|u - u_h|_{H^1(\Omega)}$ for the example considered. The convergence plot shows that an anisotropic adaptive mesh requires

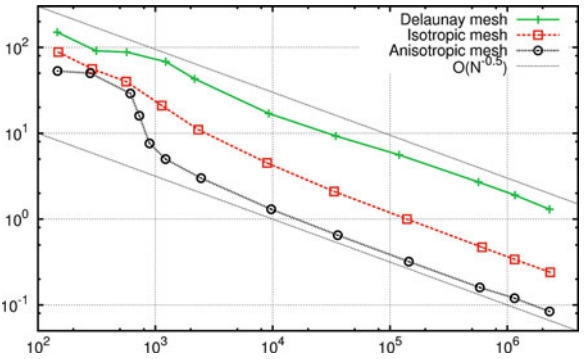


Fig. 3 Energy norm of the finite element error vs. number of mesh elements

ca. 200 times fewer elements than a quasi-uniform mesh in order to achieve the same accuracy and ca. 10 times fewer elements than an isotropic adaptive mesh. In other words, the finite element solution with an anisotropic mesh has a 15 times smaller error than an error of the solution on a quasi-uniform mesh with the same number of elements and 3 times smaller than the error achieved by means of an isotropic adaptive mesh. The asymptotic convergence order of the error in the energy norm is the same for all three kinds of meshes: it is $O(N^{-0.5})$.[2] This is expected since we cannot have a better convergence order for anisotropic mesh adaptation but can expect a much smaller constant when the solution of the problem has anisotropic features. In our test example we gain more than one order of magnitude in comparison to quasi-uniform meshes and about one half of the order in comparison to the isotropic adaptation.

Figure 3 provides also an interesting insight into the behaviour of anisotropic mesh adaptation. For very coarse meshes ($N < 300$) the resolution is not good enough to capture the anisotropy of the solution, the mesh is isotropic and has the same error as with isotropic mesh adaptation. The interesting part of the plot is between $N \approx 300$ and $N \approx 1,000$, where the algorithm starts to catch the anisotropic features and the error drops quickly. When the anisotropic mesh is fine enough to resolve the anisotropy of the solution ($N > 1,000$), the error convergence rate reaches the asymptotic state.

3.2 Condition Number of the Stiffness Matrix

In this section, we compute the exact condition number (with respect to the $\|\cdot\|_2$ matrix norm) for the stiffness matrix of the anisotropic finite elements equations and compare it to the conditioning of the finite element equations with isotropic adaptive and quasi-uniform meshes.

[2]Note that $O(N^{-0.5}) = O(h)$ for quasi-uniform meshes in 2D.

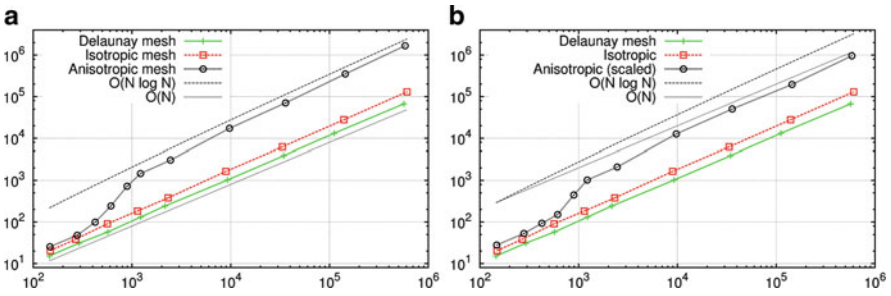


Fig. 4 Condition number of the stiffness matrix vs. number of elements. (**a**) Unscaled. (**b**) After diagonal scaling

The analysis in [22] for the Laplace operator in 2D shows that the condition number can be bounded by a term depending mainly on the number N and the largest aspect ratio of the mesh elements. In our numerical experiment the maximum aspect ratio is up to 3.8 for quasi-uniform and isotropic meshes and up to 37.9 for anisotropic meshes. Thus, the rough estimate on the ratio between the condition numbers of the anisotropic and isotropic systems should be about $37.9/3.8 \approx 10.0$. This is in perfect agreement with our numerical results presented in Fig. 4a which show that the condition number of the linear system with anisotropic meshes is about one order of magnitude higher than that with the isotropic meshes. Notice also the sudden jump in the condition number for the anisotropic case in the range $300 \leq N \leq 1,000$: the algorithm starts to catch the anisotropic features of the solution and the maximum aspect ratio of the mesh increases quickly as the mesh becomes more and more anisotropic (cp. the corresponding error decrease in Fig. 3).

Figure 4a also shows that the asymptotic behaviour of the condition number with anisotropic meshes is at most $O(N \log N)$ which is only slightly larger than $O(N)$ in the quasi-uniform case. The conditioning with isotropic adaptive meshes is also slightly larger than $O(N)$ although still smaller than $O(N \log N)$. Moreover, if a mesh is only locally anisotropic (as in our example), a proper diagonal scaling can reduce the conditioning of the stiffness matrix so that it is comparable with the condition number in the uniform case (see [22] for more details on diagonal scaling). Figure 4b shows that the asymptotic rate of the conditioning of the scaled stiffness matrix is reduced to essentially $O(N)$, which is comparable to that with uniform meshes.

4 Conclusion

Our numerical experiment shows that for problems with anisotropy the anisotropic mesh adaptation is clearly superior to the isotropic one. In our example, at least a half order of magnitude could be gained in accuracy by switching from the isotropic

mesh adaptation to the anisotropic one. The globally defined hierarchical basis error estimate provides good directional information for the anisotropic mesh generation, provided the number of mesh elements is large enough to resolve the anisotropy of the solution. It is worth pointing out that the results in Fig. 3 present the error of the final numerical solution, i.e., *after* solving the linear system. Thus, even if the condition number of the linear system with anisotropic meshes is larger than that with isotropic meshes, the accuracy gained through the anisotropic discretization for problems with anisotropic features outbalances possible losses due to the numerical accuracy.

Acknowledgements This research was supported in part by the National Science Foundation (U.S.A.) through grant DMS-1115118 and the German Research Foundation through grants SFB568/3, SPP1276 (MetStroem) and KA 3215/1-1. The authors are grateful to the anonymous referee for the valuable comments.

References

1. A. Agouzal, K. Lipnikov, and Y. Vassilevski. Generation of quasi-optimal meshes based on a posteriori error estimates. In *Proceedings of the 16th International Meshing Roundtable*, pages 139–148, 2008.
2. A. Agouzal, K. Lipnikov, and Y. Vassilevski. Anisotropic mesh adaptation for solution of finite element problems using hierarchical edge-based error estimates. In *Proceedings of the 18th International Meshing Roundtable*, pages 595–610, 2009.
3. A. Agouzal, K. Lipnikov, and Y. Vassilevski. Hessian-free metric-based mesh adaptation via geometry of interpolation error. *Comput. Math. Math. Phys.*, 50(1):124–138, Jan. 2010.
4. T. Apel, S. Grosman, P. K. Jimack, and A. Meyer. A new methodology for anisotropic mesh refinement based upon error gradients. *Appl. Numer. Math.*, 50(3–4):329–341, 2004.
5. W. Cao, W. Huang, and R. D. Russell. Comparison of two-dimensional r-adaptive finite element methods using various error indicators. *Math. Comput. Simulation*, 56(2):127–143, 2001.
6. E. F. D'Azevedo. Optimal triangular mesh generation by coordinate transformation. *SIAM J. Sci. Stat. Comput.*, 12(4):755–786, 1991.
7. E. F. D'Azevedo, C. H. Romine, and J. M. Donato. Coefficient adaptive trianglation for strongly anisotroic problems. Technical Report ORNL/TM-13086, Oak Ridge National Laboratory, 1997.
8. M. Dobrowolski, S. Gräf, and C. Pflaum. On a posteriori error estimators in the finite element method on anisotropic meshes. *Electron. Trans. Numer. Anal.*, 8:36–45, 1999.
9. V. Dolejší. Anisotropic mesh adaptation for finite volume and finite element methods on triangular meshes. *Comput. Vis. Sci.*, 1(3):165–178, 1998.
10. L. Formaggia, S. Micheletti, and S. Perotto. Anisotropic mesh adaptation in computational fluid dynamics: Application to the advection-diffusion-reaction and the stokes problems. *Appl. Numer. Math.*, 51(4):511–533, 2004. Applied Scientific Computing: Advances in Grid Generatuion, Approximation and Numerical Modeling.
11. L. Formaggia and S. Perotto. New anisotropic a priori error estimates. *Numer. Math.*, 89(4):641–667, 2001.
12. L. Formaggia and S. Perotto. Anisotropic error estimates for elliptic problems. *Numer. Math.*, 94(1):67–92, 2003.

13. J. Fröhlich, J. Lang, and R. Roitzsch. Selfadaptive finite element computations with smooth time controller and anisotropic refinement. In *Numerical Methods in Engineering '96*, pages 523–527. John Wiley & Sons, New York, 1996.
14. E. H. Georgoulis, E. Hall, and P. Houston. Discontinuous Galerkin methods for advection-diffusion-reaction problems on anisotropically refined meshes. *SIAM J. Sci. Comput.*, 30(1):246–271, 2007.
15. S. Grosman. *Adaptivity in anisotriopic finite element calculations*. PhD thesis, Technische Universität München, 2006.
16. F. Hecht. *BAMG*. http://www.ann.jussieu.fr/hecht/ftp/bamg/.
17. W. Huang. Metric tensors for anisotropic mesh generation. *J. Comput. Phys.*, 204(2):633–665, 2005.
18. W. Huang. Mathematical principles of anisotropic mesh adaptation. *Commun. Comput. Phys.*, 1(2):276–310, 2006.
19. W. Huang, L. Kamenski, and J. Lang. A new anisotropic mesh adaptation method based upon hierarchical a posteriori error estimates. *J. Comput. Phys.*, 229(6):2179–2198, 2010.
20. W. Huang and X. Li. An anisotropic mesh adaptation method for the finite element solution of variational problems. *Finite Elem. Anal. Des.*, 46(1–2):61–73, 2010.
21. L. Kamenski. *Anisotropic Mesh Adaptation Based on Hessian Recovery and A Posteriori Error Estimates*. PhD thesis, TU Darmstadt, 2009.
22. L. Kamenski, W. Huang, and H. Xu. Conditioning of finite element equations with arbitrary anisotropic meshes. Submitted, e-print: arXiv:1201.3651, 2012.
23. X. Li and W. Huang. An anisotropic mesh adaptation method for the finite element solution of heterogeneous anisotropic diffusion problems. *J. Comput. Phys.*, 229(21):8072–8094, 2010.
24. W. F. Mitchell. A collection of 2d elliptic problems for testing adaptive algorithms. Technical Report NISTIR 7668, National Institute of Standards and Technology, 2010.
25. M. Picasso. An anisotropic error indicator based on Zienkiewicz–Zhu error estimator: Application to elliptic and parabolic problems. *SIAM J. Sci. Comput.*, 24(4):1328–1355, 2003.
26. M. Picasso. Adaptive finite elements with large aspect ratio based on an anisotropic error estimator involving first order derivatives. *Comput. Methods Appl. Mech. Engrg.*, 196(1–3):14–23, 2006.
27. R. B. Simpson. Anisotropic mesh transformations and optimal error control. *Appl. Numer. Math.*, 14(1–3):183 – 198, 1994.
28. Y. Vassilevski and K. Lipnikov. An adaptive algorithm for quasioptimal mesh generation. *Comput. Math. Math. Phys.*, 39(9):1468–1486, 1999.

Anisotropic Recovery-Based a Posteriori Error Estimators for Advection-Diffusion-Reaction Problems

S. Micheletti and S. Perotto

Abstract We combine the good properties of recovery-based error estimators with the richness of information typical of an anisotropic a posteriori analysis. This merging yields error estimators which are general purpose yet simple and easy to implement, and automatically incorporate detailed geometric information about the computational mesh. This allows us to devise an effective anisotropic mesh adaptation procedure suited to control the discretization error both in the energy norm and in a goal-oriented framework. The advection-diffusion-reaction problem is considered as a computational paradigm.

1 Introduction

Advection-diffusion-reaction (ADR) problems can be interesting per se (e.g., pollution transport in air or rivers, population dynamics in biology) or can be employed as downscaled models for studying more complex problems in computational fluid dynamics (e.g., the Navier-Stokes equations for modeling viscous flows around bodies). A joint effect of geometry, advective field pattern, and boundary conditions may sometimes render ADR problems hard to be numerically solved unless ad hoc numerical schemes or computational meshes are employed.

The objective of this work is to propose practical a posteriori error estimators for driving an anisotropic adaptation of the mesh, i.e., where not only the size but also the shape and the orientation of the elements are controlled so as to match the directional features of the solution. It is in fact well known that anisotropic mesh adaptation is cost-effective in dealing with a broad range of problems [2–4, 8].

S. Micheletti (✉) · S. Perotto
MOX – Dipartimento di Matematica "F. Brioschi", Politecnico di Milano, Piazza Leonardo da Vinci 32, I-20133, Milano (MI), Italy
e-mail: stefano.micheletti@polimi.it; simona.perotto@polimi.it

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_5,

In particular, we stick to recovery-based estimators, relying on the ideas proposed by Zienkiewicz and Zhu in [9]. After devising a simple recovery technique, different from the standard one, we introduce an estimator for controlling the H^1-seminorm of the discretization error [5–7]. This estimator automatically includes the anisotropic features (size, aspect ratio, and orientation) of the triangulation in contrast to the standard Zienkiewicz-Zhu estimator.

The strong interest in approximating goal quantities for practical applications led us to extend the theory in [5–7] to a goal-oriented setting, showing that recovery-based and goal-oriented are compatible approaches which can be merged in an effective and practical way.

As a reference ADR problem used to introduce the new anisotropic estimator we employ the standard one completed with homogeneous Dirichlet boundary conditions, i.e., find $u \in V = H_0^1(\Omega)$, such that

$$a(u,v) = \int_\Omega \mu\, \nabla u \cdot \nabla v \, d\mathbf{x} + \int_\Omega \boldsymbol{\beta} \cdot \nabla u \, v \, d\mathbf{x} + \int_\Omega \sigma \, u \, v \, d\mathbf{x} = \int_\Omega f \, v \, d\mathbf{x} \quad \forall v \in V, \tag{1}$$

where Ω is a polygonal domain in $\mathbb{R}^2$, $\mu > 0$ is the diffusion coefficient, $\boldsymbol{\beta} \in [W^{1,\infty}(\Omega)]^2$ is the advective field, $\sigma \in L^\infty(\Omega)$ is the reactive coefficient, and where standard notation are adopted for the Lebesgue and Sobolev spaces and their norms. To guarantee the well-posedness of (1) we add the assumption $\sigma - \frac{1}{2}\nabla \cdot \boldsymbol{\beta} \geq 0$.

The structure of a recovery-based estimator allows us a straightforward extension of the a posteriori analysis below to other types of boundary conditions.

2 Zienkiewicz-Zhu Like Anisotropic Error Estimators

The good properties of the recovery-based error estimators (independence of the problem, computational easiness, effectiveness) justify their broad use, mostly because they work pretty well in practice in many engineering applications. On the other hand, the presence of strong directional features in such applications requires ad hoc meshes to sharply detect the phenomena of interest. To meet this demand, we have proposed in [5–7] a suitable enrichment of the standard recovery-based estimators, which explicitly takes into account the intrinsic directionalities of the problem. For this purpose, let us first lay down the anisotropic background.

2.1 The Anisotropic Setting

Let $\mathcal{T}_h = \{K\}$ be a conforming partition of Ω consisting of triangles. According to the anisotropic framework in [1], the size, shape, and orientation of each element

K of $\mathscr{T}_h$ are described by means of the affine map $T_K : \widehat{K} \to K$ between the reference triangle $\widehat{K}$ and the generic element $K \in \mathscr{T}_h$. In particular, we pick $\widehat{K}$ as the equilateral triangle centered at the origin, with coordinates $(-\sqrt{3}/2, -1/2)$, $(\sqrt{3}/2, -1/2)$, $(0, 1)$ and edge length $\sqrt{3}$. The map T_K writes out as $\mathbf{x} = T_K(\widehat{\mathbf{x}}) = M_K\,\widehat{\mathbf{x}} + \mathbf{t}_K$, where $M_K \in \mathbb{R}^{2\times 2}$ is the Jacobian and $\mathbf{t}_K \in \mathbb{R}^2$ is the shift vector. To get the anisotropic information associated with K out of T_K, we factorize M_K via the polar decomposition as $M_K = B_K\,Z_K$, where $B_K \in \mathbb{R}^{2\times 2}$ is symmetric positive definite, and $Z_K \in \mathbb{R}^{2\times 2}$ is orthogonal. Then B_K is spectrally decomposed as $B_K = R_K^T\,\Lambda_K\,R_K$, with $R_K^T = [\mathbf{r}_{1,K}, \mathbf{r}_{2,K}]$ and $\Lambda_K = \mathrm{diag}(\lambda_{1,K}, \lambda_{2,K})$ the eigenvector and eigenvalue matrix, respectively. The map T_K stretches the unit circle circumscribing $\widehat{K}$, into an ellipse circumscribing K: the unit vectors $\{\mathbf{r}_{i,K}\}$ provide us with the corresponding principal directions, whereas the eigenvalues $\{\lambda_{i,K}\}$ are the length of the ellipse semi-axes. Without loss of generality, we assume $\lambda_{1,K} \geq \lambda_{2,K} > 0$ so that the aspect ratio, $s_K = \lambda_{1,K}/\lambda_{2,K}$ is always greater than or equal to one, for any $K \in \mathscr{T}_h$, equality holding when K is equilateral.

2.2 An Error Estimator for the H^1-Seminorm

In [5] we propose an a posteriori error estimator for the H^1-seminorm of the discretization error $e_h = u - u_h$, where u_h is the Galerkin affine finite element approximation to (1). The actual estimator reads

$$\eta_{H^1}^2 = \sum_{K\in\mathscr{T}_h} \left[\eta_{K,H^1}\right]^2, \quad \left[\eta_{K,H^1}\right]^2 = \frac{1}{\lambda_{1,K}\lambda_{2,K}} \sum_{i=1}^{2} \lambda_{i,K}^2 \left(\mathbf{r}_{i,K}^T\, G_K(\mathbf{E}_K(u_h))\mathbf{r}_{i,K}\right), \tag{2}$$

where $G_K(\cdot)$ is the symmetric positive semidefinite matrix with entries

$$[G_K(\mathbf{w})]_{i,j} = \sum_{T\in\Delta_K} \int_T w_i\, w_j\, d\mathbf{x}, \quad \text{with } i, j = 1, 2, \tag{3}$$

with $\Delta_K = \{T \in \mathscr{T}_h : T \cap K \neq \emptyset\}$, and where $\mathbf{E}_K(u_h) = P_{\Delta_K}(\nabla u_h) - \nabla u_h|_{\Delta_K}$ is the approximation, over Δ_K, to the error on the gradient via a suitable recovered gradient $P_{\Delta_K}(\nabla u_h)$ [9]. In particular, in [5–7] we employ as recovery procedure the area-weighted average over the patch Δ_K of the gradients of the discrete solution. Estimator (2) exhibits the standard recovery-based structure in the term $\mathbf{E}_K(u_h)$, while the anisotropic contribution is represented by the weighted projection of the isotropic estimator onto the anisotropic directions $\mathbf{r}_{i,K}$. This projection is the novelty with respect to [9], and allows one to steer a proper anisotropic adaptation.

2.3 A Goal-Oriented Error Estimator

The strong interest in engineering applications prompted us in [7] to generalize estimator (2) to a goal-oriented approach. That approach is however constrained to the Poisson problem and to a special choice of the functional of interest $J : V \to \mathbb{R}$. Here we propose a more general approach suited to dealing with problem (1) and where J can be any functional in the dual space V'. The dual problem associated with (1) is: find $z \in V$, such that

$$a(v, z) = J(v) \quad \forall v \in V. \tag{4}$$

Combining (1) with (4) and using the Galerkin orthogonality, we get the error representation

$$J(u - u_h) = a(u - u_h, z - z_h), \tag{5}$$

with z_h the Galerkin affine finite element approximation to (4). In a recovery-based spirit, (5) suggests the quantity

$$\int_{\Delta_K} \mu\, \mathbf{E}_K(u_h) \cdot \mathbf{E}_K(z_h)\, d\mathbf{x} + \int_{\Delta_K} \boldsymbol{\beta} \cdot \mathbf{E}_K(u_h) F_K(z_h)\, d\mathbf{x} + \int_{\Delta_K} \sigma\, F_K(u_h) F_K(z_h)\, d\mathbf{x} \tag{6}$$

as a first attempt to estimate $J(e_h)\big|_K$, where the explicit definition of $a(\cdot,\cdot)$ is used and suitable recovered quantities replace the exact fields. In particular, $F_K(u_h) = \big(R(u_h) - u_h\big)\big|_{\Delta_K}$, where $R(u_h)$ is the affine field recovered via the arithmetic average over the patch $\Delta_N = \{T \in \mathcal{T}_h : T \ni N\}$ of u_h at the centroids of $T \in \Delta_N$, with N the generic node of $\mathcal{T}_h$.

The next step is to convert (6) into an anisotropic source of information. The strategy that we pursue casts the first three integrals of (1) in the reference framework ($\widehat{\Delta}_K = T_K^{-1}(\Delta_K)$) and then carries them back to the physical domain, employing the spectral properties of T_K. This leads for free to a structure similar to the one in (2), i.e., with built-in anisotropic quantities. Let us exemplify this procedure starting from the diffusive term. We employ the relations $\widehat{\nabla}\widehat{u} = M_K^T \nabla u$, $|\widehat{\Delta}_K| = |\Delta_K|/(\lambda_{1,K}\lambda_{2,K})$, and $\widehat{u} = u \circ T_K$ (and similarly for v), and the decompositions of the Jacobian M_K in Sect. 2.1, to get

$$\begin{aligned}
&\int_{\widehat{\Delta}_K} \widehat{\mu}\widehat{\nabla}\widehat{u} \cdot \widehat{\nabla}\widehat{v}\, d\widehat{\mathbf{x}} = \frac{1}{\lambda_{1,K}\lambda_{2,K}} \int_{\Delta_K} \mu\, \Lambda_K R_K(\nabla u) \cdot \Lambda_K R_K \nabla v\, d\mathbf{x} \\
&= \int_{\Delta_K} \mu \big[s_K (\mathbf{r}_{1,K} \cdot \nabla u)(\mathbf{r}_{1,K} \cdot \nabla v) + s_K^{-1} (\mathbf{r}_{2,K} \cdot \nabla u)(\mathbf{r}_{2,K} \cdot \nabla v) \big]\, d\mathbf{x} \\
&= s_K \mathbf{r}_{1,K}^T G_{K,\mu}(\nabla u, \nabla v) \mathbf{r}_{1,K} + s_K^{-1} \mathbf{r}_{2,K}^T G_{K,\mu}(\nabla u, \nabla v) \mathbf{r}_{2,K},
\end{aligned} \tag{7}$$

where $G_{K,\mu}$ is the matrix with entries $[G_{K,\mu}(\mathbf{t},\mathbf{w})]_{ij} = \int_{\Delta_K} \mu \mathbf{t}_i \mathbf{w}_j \, d\mathbf{x}$, $i, j = 1, 2$, for $\mathbf{t}, \mathbf{w} : \Omega \to \mathbb{R}^2$. The consistency with the isotropic case is guaranteed, i.e., if $\lambda_{1,K} = \lambda_{2,K}$, (7) coincides with $\int_{\Delta_K} \mu \nabla u \cdot \nabla v \, d\mathbf{x}$, up to a scaling factor.

In an analogous manner, the advective term becomes

$$
\begin{aligned}
&(\lambda_{1,K}\lambda_{2,K})^{1/2} \int_{\widehat{\Delta}_K} \widehat{\boldsymbol{\beta}} \cdot \widehat{\nabla}\widehat{u}\, \widehat{v} \, d\widehat{\mathbf{x}} = (\lambda_{1,K}\lambda_{2,K})^{-1/2} \int_{\Delta_K} \boldsymbol{\beta}^T Z_K^T R_K^T \Lambda_K R_K \nabla u \, v \, d\mathbf{x} \\
&= (\lambda_{1,K}\lambda_{2,K})^{-1/2} \int_{\Delta_K} (Z_K \boldsymbol{\beta})^T \left[\lambda_{1,K} (\mathbf{r}_{1,K} \cdot \nabla u) \mathbf{r}_{1,K} + \lambda_{2,K} (\mathbf{r}_{2,K} \cdot \nabla u) \mathbf{r}_{2,K} \right] v \, d\mathbf{x} \\
&= s_K^{1/2} \mathbf{r}_{1,K}^T G_{K,\boldsymbol{\beta}}(\nabla u, v) \mathbf{r}_{1,K} + s_K^{-1/2} \mathbf{r}_{2,K}^T G_{K,\boldsymbol{\beta}}(\nabla u, v) \mathbf{r}_{2,K}, \qquad (8)
\end{aligned}
$$

where the entries of $G_{K,\boldsymbol{\beta}}$ are $[G_{K,\boldsymbol{\beta}}(\mathbf{t}, w)]_{ij} = \int_{\Delta_K} (Z_K \boldsymbol{\beta})_i \mathbf{t}_j w \, d\mathbf{x}$, $i, j = 1, 2$, for $\mathbf{t} : \Omega \to \mathbb{R}^2$ and $w : \Omega \to \mathbb{R}$. The consistency with the isotropic case is recovered via the scaling factor $(\lambda_{1,K}\lambda_{2,K})^{1/2}$.

The reactive term does not provide any anisotropic contribution.

The right-hand sides in (7) and (8) yield the anisotropic counterpart of the first two terms in (6) after replacing ∇u with $\mathbf{E}_K(u_h)$, ∇v with $\mathbf{E}_K(z_h)$, and ∇u with $\mathbf{E}_K(u_h)$, v with $F_K(z_h)$, respectively. This suggests as a first anisotropic attempt to estimate $J(e_h)\big|_K$ the quantity

$$
\begin{aligned}
&s_K \mathbf{r}_{1,K}^T G_{K,\mu}(\mathbf{E}_K(u_h), \mathbf{E}_K(z_h)) \mathbf{r}_{1,K} + s_K^{-1} \mathbf{r}_{2,K}^T G_{K,\mu}(\mathbf{E}_K(u_h), \mathbf{E}_K(z_h)) \mathbf{r}_{2,K} \\
&+ s_K^{1/2} \mathbf{r}_{1,K}^T G_{K,\boldsymbol{\beta}}(\mathbf{E}_K(u_h), F_K(z_h)) \mathbf{r}_{1,K} + s_K^{-1/2} \mathbf{r}_{2,K}^T G_{K,\boldsymbol{\beta}}(\mathbf{E}_K(u_h), F_K(z_h)) \mathbf{r}_{2,K} \\
&+ \int_{\Delta_K} \sigma \, F_K(u_h) F_K(z_h) \, d\mathbf{x}. \qquad (9)
\end{aligned}
$$

To make such an estimator effective, it could be practical to deal with symmetric positive definite matrices, which is not the case of $G_{K,\mu}$ and $G_{K,\boldsymbol{\beta}}$. Consequently, since

$$
\begin{aligned}
\mathbf{r}_{i,K}^T G_{K,\mu}(\mathbf{t}, \mathbf{w}) \mathbf{r}_{i,K} &= \mathbf{r}_{i,K}^T G_{K,\mu}(\mathbf{w}, \mathbf{t}) \mathbf{r}_{i,K} = \mathbf{r}_{i,K}^T G_{K,\mu}^T(\mathbf{t}, \mathbf{w}) \mathbf{r}_{i,K}, \\
\mathbf{r}_{i,K}^T G_{K,\boldsymbol{\beta}}(\mathbf{t}, w) \mathbf{r}_{i,K} &= \mathbf{r}_{i,K}^T G_{K,Z_K^T \mathbf{t}}(Z_k \boldsymbol{\beta}, w) \mathbf{r}_{i,K} = \mathbf{r}_{i,K}^T G_{K,\boldsymbol{\beta}}^T(\mathbf{t}, w) \mathbf{r}_{i,K},
\end{aligned}
$$

$i = 1, 2$, we can replace in (9) the two matrices with their symmetric counterparts $G_{K,\mu}^{sym}(\cdot,\cdot) = (G_{K,\mu}(\cdot,\cdot) + G_{K,\mu}^T(\cdot,\cdot))/2$ and $G_{K,\boldsymbol{\beta}}^{sym}(\cdot,\cdot) = (G_{K,\boldsymbol{\beta}}(\cdot,\cdot) + G_{K,\boldsymbol{\beta}}^T(\cdot,\cdot))/2$.

Next, to ensure the positive definiteness, we replace the symmetrized matrices with the modulus matrices (e.g., if $G = V^T D V$, then $|G| = V^T |D| V$, with D and V the eigenvalues and eigenvectors matrices, respectively); this leads to the definitive estimator

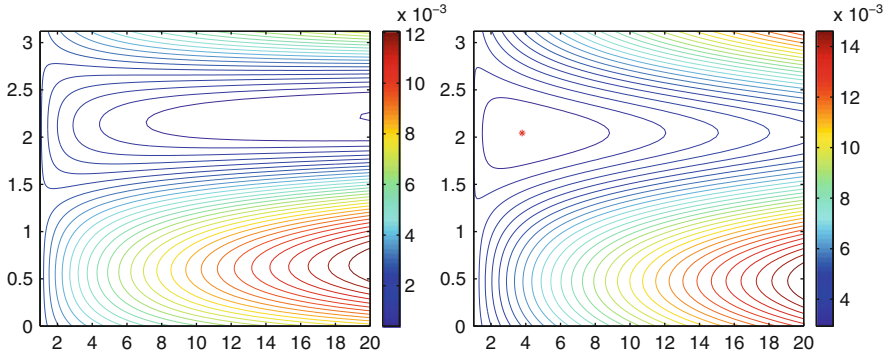


Fig. 1 Contour lines of the absolute value of (9) (*left*) and of $\eta_{K,J}$ (*right*): the *star* marks the minimum

$$\begin{aligned}\eta_{K,J} &= s_K \mathbf{r}_{1,K}^T \left| G_{K,\mu}^{sym}(\mathbf{E}_K(u_h), \mathbf{E}_K(z_h)) \right| \mathbf{r}_{1,K} + s_K^{-1} \mathbf{r}_{2,K}^T \left| G_{K,\mu}^{sym}(\mathbf{E}_K(u_h), \mathbf{E}_K(z_h)) \right| \mathbf{r}_{2,K} \\ &+ s_K^{1/2} \mathbf{r}_{1,K}^T \left| G_{K,\boldsymbol{\beta}}^{sym}(\mathbf{E}_K(u_h), F_K(z_h)) \right| \mathbf{r}_{1,K} + s_K^{-1/2} \mathbf{r}_{2,K}^T \left| G_{K,\boldsymbol{\beta}}^{sym}(\mathbf{E}_K(u_h), F_K(z_h)) \right| \mathbf{r}_{2,K} \\ &+ \left| \int_{\Delta_K} \sigma \, F_K(u_h) F_K(z_h) \, d\mathbf{x} \right|.\end{aligned}$$

An example of the benefits due to the regularization above is shown in Fig. 1, where we compare the contour lines associated with the absolute value of the quantity in (9) with those of $\eta_{K,J}$, for given indefinite matrices $G_{K,\mu}$, $G_{K,\boldsymbol{\beta}}$ associated with the test case in Sect. 3. Only in the second case we get a unique minimum.

Finally, the matrix Z_K in the definition of $G_{K,\boldsymbol{\beta}}(\cdot,\cdot)$ is, in practice, taken as the identity matrix. It represents a degree of freedom in the mesh generation associated with a rotation of K inside the ellipse given by $\{\lambda_{i,K}, \mathbf{r}_{i,K}\}_{i=1,2}$. However this information is, usually, not required by a metric-based mesh generator.

3 Numerical Assessment

We provide here the actual iterative procedure employed to convert $\eta_{K,J}$ into a practical tool for driving the mesh adaptation. The idea is to compute u_h and z_h on the actual grid; then, via $\eta_{K,J}$, we predict the new adapted grid. Thus, at each iteration, the unknown is the adapted grid.

3.1 The Adaptive Procedure

Following, e.g., [7], we first properly scale the matrices in $\eta_{K,J}$ with respect to $|\Delta_K|$, to factor out the patch size information. This yields

$$\eta_{K,J} = |\widehat{\Delta}_K|\lambda_{1,K}\lambda_{2,K}\Big[s_K\mathbf{r}_{1,K}^T\big|\widetilde{G}_{K,\mu}^{sym}(\mathbf{E}_K(u_h),\mathbf{E}_K(z_h))\big|\mathbf{r}_{1,K}$$
$$+ s_K^{-1}\mathbf{r}_{2,K}^T\big|\widetilde{G}_{K,\mu}^{sym}(\mathbf{E}_K(u_h),\mathbf{E}_K(z_h))\big|\mathbf{r}_{2,K} + s_K^{1/2}\mathbf{r}_{1,K}^T\big|\widetilde{G}_{K,\boldsymbol{\beta}}^{sym}(\mathbf{E}_K(u_h),F_K(z_h))\big|\mathbf{r}_{1,K}$$
$$+ s_K^{-1/2}\mathbf{r}_{2,K}^T\big|\widetilde{G}_{K,\boldsymbol{\beta}}^{sym}(\mathbf{E}_K(u_h),F_K(z_h))\big|\mathbf{r}_{2,K} + \frac{1}{|\Delta_K|}\Big|\int_{\Delta_K}\sigma\, F_K(u_h)F_K(z_h)\,d\mathbf{x}\Big|\Big],$$

where $\widetilde{G}_{K,\mu}^{sym}(\cdot,\cdot) = G_{K,\mu}^{sym}(\cdot,\cdot)/|\Delta_K|$, and likewise for $\widetilde{G}_{K,\boldsymbol{\beta}}^{sym}(\cdot,\cdot)$. In a *predictive* fashion, these matrices and the area $|\Delta_K|$ are computed on the actual mesh, whereas $\lambda_{i,K}, \mathbf{r}_{i,K}$, $i = 1,2$, become the quantities to be predicted. For this purpose, we minimize the expression in square brackets with respect to the pair $\{s_K, \mathbf{r}_{1,K}\}$ subject to the constraints $s_K \geq 1$ and $\mathbf{r}_{1,K}\cdot\mathbf{r}_{2,K} = 0$, with $\|\mathbf{r}_{1,K}\| = \|\mathbf{r}_{2,K}\| = 1$. With this aim, we set $\mathbf{r}_{1,K} = [\cos\theta, \sin\theta]^T$ and $\mathbf{r}_{2,K} = [-\sin\theta, \cos\theta]^T$, for a certain $0 \leq \theta < \pi$, and let $F = F(s_K, \theta)$ be the quantity in brackets. For this minimization we use the Matlab function `fmincon`. This yields the minimum $F^* = F(s_K^*, \theta^*)$ for the optimal values, $\{s_K^*, \theta^*\}$, and consequently $\mathbf{r}_{1,K}^* = [\cos\theta^*, \sin\theta^*]^T$.

To get the optimal values $\lambda_{1,K}^*$ and $\lambda_{2,K}^*$, we enforce the equidistribution of the error, i.e., $\eta_{K,J} = \texttt{TOLL}/\#\mathscr{T}_h$, where $\#\mathscr{T}_h$ is the mesh cardinality and TOLL is the accuracy demanded on $J(e_h)$. This yields $\lambda_{1,K}^*\lambda_{2,K}^* = \texttt{TOLL}/(\#\mathscr{T}_h|\widehat{\Delta}_K|F^*)$. To split the values $\lambda_{1,K}^*, \lambda_{2,K}^*$ we finally exploit the identity $s_K^* = \lambda_{1,K}^*/\lambda_{2,K}^*$.

The optimal metric is formed by the optimal values $\lambda_{1,K}^*, \lambda_{2,K}^*, \mathbf{r}_{1,K}^*$.

3.2 The "Arrow" Test Case

We consider (1), choosing $\mu = \sigma = 10^{-2}$, $\boldsymbol{\beta} = [1,1]^T$ on $\Omega = (0,1)^2$. The exact solution is tailor-made so that it exhibits one internal layer along the SW-NE diagonal and two boundary layers along the top and right sides of Ω (see Fig. 2 (left)):

$$u(x,y) = \big[\alpha(x,y) + \rho(x)\,\rho(y)\big]\delta(x)\delta(y),$$

with $\alpha(x,y) = e^{-(y-x)^2/0.01}$, $\rho(\zeta) = \zeta - \big(e^{(\zeta-1)/\varepsilon} - e^{-1/\varepsilon}\big)/\big(1 - e^{-1/\varepsilon}\big)$, $\delta(\zeta) = 1 - e^{-\zeta/\varepsilon} + e^{-1/\varepsilon} - e^{-(1-\zeta)/\varepsilon}$, and $\varepsilon = 10^{-2}$. The source term is computed as $f = -\mu\,\Delta u + \boldsymbol{\beta}\cdot\nabla u + \sigma\,u$.

On this test case we assess the performance of both the estimators defined by η_{K,H^1} and $\eta_{K,J}$. Let us start from the H^1-seminorm control. In Fig. 3 (left) we show the adapted grid for the tolerance $\texttt{TOLL} = 10^{-2}/2$: it consists of 3,887 elements which perfectly capture all the internal and boundary layers. The maximum stretching factor over the mesh elements is $s_K^{\max} = 129.1$. The convergence history for this estimator is summarized in Fig. 2 (center) as a function of $\#\mathscr{T}_h$: the rate of convergence turns out to be about $1/2$, accordingly to the a priori analysis.

Moving to the goal-oriented setting, we consider two different goal-functionals: we control the mean value of e_h on Ω via J_1 and the energy norm $a(e_h,e_h)$ of e_h

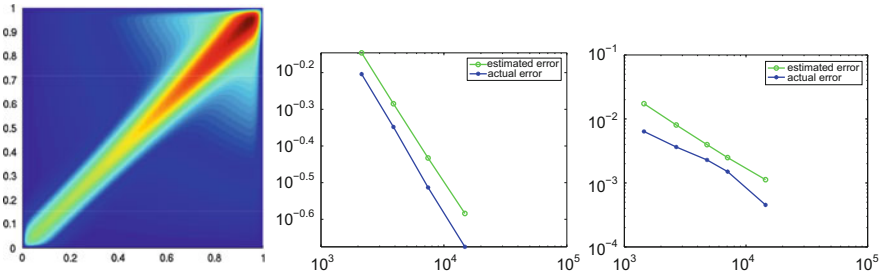


Fig. 2 Solution u (*left*) and convergence history for the H^1-seminorm (*center*) and for $J = J_1$ (*right*)

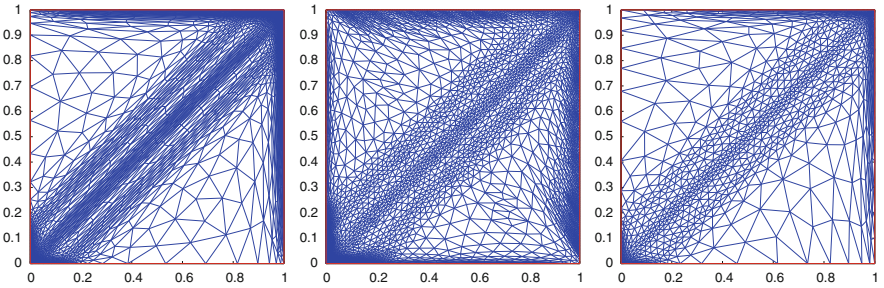


Fig. 3 Adapted grids for the H^1-seminorm (*left*), $J = J_1$ (*center*) and $J = J_2$ (*right*) control

via J_2.[1] The grids associated with these two choices are quite different (see Fig. 3, (center) and (right)): in the case of J_1 we can appreciate the strong influence of the dual problem through the boundary layers on the left and bottom sides of Ω; this is not the case for J_2 since the control of the energy norm leads to identifying z with u. Moreover, the directions of the anisotropic features on the top and right sides are skew and parallel to these layers for $J = J_1$ and $J = J_2$, respectively. These differences confirm the sensitivity of the adapted mesh to the goal functional. The maximum stretching factor and the cardinality of $\mathcal{T}_h$ are $s_K^{\max} = 29.9$, $\#\mathcal{T}_h = 7,061$, and $s_K^{\max} = 38.4$, $\#\mathcal{T}_h = 2,103$ in the two cases, for $\texttt{TOLL} = 10^{-2}/4$ and $\texttt{TOLL} = 10^{-1}/2$, respectively.

Figure 2 (right) displays the convergence history for the functional error $J_1(e_h)$ which exhibits an $O(1/\#\mathcal{T}_h)$ order of convergence.

Prompted us by the above promising results, we are now extending the approach proposed in this paper to the more challenging shallow water system.

[1] Indeed, picking $J_2(\varphi) = a(\varphi, u)$, we get that $J_2(e_h) = a(e_h, u) = a(e_h, e_h)$, thanks to the Galerkin orthogonality.

References

1. Formaggia, L., Perotto, S.: New anisotropic a priori error estimates. Numer. Math. **89**(4), 641–667 (2001)
2. Frey, P.J., Alauzet, F.: Anisotropic mesh adaptation for CFD computations. Comput. Methods Appl. Mech. Engrg. **194**, 5068–5082 (2005)
3. Gruau, C., Coupez, T.: 3D tetrahedral, unstructured and anisotropic mesh generation with adaptation to natural and multidomain metric. Comput. Methods Appl. Mech. Engrg. **194** (48–49), 4951–4976 (2005)
4. Micheletti, S., Perotto, S.: Output functional control for nonlinear equations driven by anisotropic mesh adaption: The Navier-Stokes equations. SIAM J. Sci. Comput. **30** (6), 2817–2854 (2008)
5. Micheletti, S., Perotto, S.: Anisotropic adaptation via a Zienkiewicz-Zhu error estimator for 2D elliptic problems. In: Kreiss, G., Lötstedt, P., Målqvist, A., Neytcheva, M. (eds.) Numerical Mathematics and Advanced Applications, pp. 645–653. Springer-Verlag, Berlin (2010)
6. Farrell, P.E., Micheletti, S., Perotto, S.: An anisotropic Zienkiewicz-Zhu type error estimator for 3D applications. Int. J. Numer. Meth. Engng **85** (6), 671–692 (2010)
7. Farrell, P.E., Micheletti, S., Perotto, S.: A recovery-based error estimator for anisotropic mesh adaptation in CFD. Bol. Soc. Esp. Mat. Apl. **50**, 115–137 (2010)
8. Piggott, M.D., Pain, C.C., Gorman, G.J., Power, P.W., Goddard, A.J.H.: h, r, and hr adaptivity with applications in numerical ocean modelling. Ocean Model. **10** (1–2), 95–113 (2005)
9. Zienkiewicz, O.C., Zhu, J.Z.: A simple error estimator and adaptive procedure for practical engineering analysis. Int. J. Numer. Meth. Engng **24**, 337–357 (1987)

On Computable Bounds of Modeling Errors

S. Repin and T. Samrowski

Abstract We give an overview of results related to computable and guaranteed bounds of modeling errors derived with the help of a posteriori estimates of the functional type and discuss estimates of errors arising in dimension reduction, defeaturing (simplification) of highly structured models, and homogenization.

1 Introduction

One of the main questions arising in real life computations is how to estimate errors encompassed in a mathematical model. In this paper, we give an overview of the results related to guaranteed and computable estimates of modeling errors obtained with the help of mathematical techniques developed in [9, 10, 12, 18]. In these and some other publications, computable estimates of the distance to exact solutions of various boundary value problems has been derived. The respective estimates has been derived by general methods of the calculus of variations and functional analysis without attracting special properties of approximations (as, e.g., Galerkin orthogonality or superconvergence). For this reason, they are applicable for any function from the energy functional class of the problem considered. In particular, they can be applied to solutions of simplified mathematical problems arising in

- Dimension reduction models,
- Defeaturing (simplification) of models,
- Homogenization theory.

S. Repin (✉)
V.A. Steklov Institute of Mathematics, Fontanka 27, 191 011 St. Petersburg, Russia
e-mail: repin@pdmi.ras.ru

T. Samrowski
Institute of Mathematics, Zurich University, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland
e-mail: tatiana.samrowski@math.uzh.ch

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_6,

In this note, we give an overview of theoretical results obtained for the above mentioned problems. The corresponding numerical technology and results are discussed in [14, 15, 18]. Evaluation of errors generated by data uncertainty is a close subject, which can be studied by the same method (see [8, 12]).

2 Dimension Reduction Models

The theory of thin–walled constructions in solid mechanics was the one that stimulated the earliest analysis of such models connected with names of Kirchhoff, Love, Timoshenko and others. Estimation of modeling errors in dimension reduction (DR) models (by methods different from that we discuss here) were investigated by several authors see, e.g., [1, 3, 4, 20].

Classical analysis of the thin–walled structures was mainly concentrated on proving that if the thickness $d \to 0$, then a sequence of 3D solutions tends (in a certain sense) to a limit function, which can be found as a solution of a "reduced" 2D problem.

By contrast, we discuss computable and guaranteed estimates of the difference between the solutions of a 3D problem and dimensionally reduced (e.g., 2D) simplified problem.

2.1 Linear Elliptic Equations

Let us consider a 3D elastic body occupying the domain $\Omega = \widehat{\Omega} \times (-d, +d)$, $\widehat{\Omega} \subseteq \mathbb{R}^2$ with boundary $\widehat{\Gamma}$ (henceforth, variables and functions depending only on (x_1, x_2) are denoted by "hats", e.g., $\widehat{x} \in \widehat{\Omega}$). Assume that the domain is thin, i.e., $d \ll \operatorname{diam}(\widehat{\Omega}) := \sup_{(x_1,x_2)\in\widehat{\Omega}} |x_1 - x_2|$ and consider the problem

$$-\nabla \cdot (A\nabla u) + \rho^2 u = f \text{ in } \Omega\,,\ A\nabla u \cdot \nu_N = F_N \text{ on } \Gamma_N\,,\ u = 0 \text{ on } \Gamma_D\,, \tag{1}$$

where $f \in L^2(\Omega)$, $F_N \in L^2(\Gamma_N)$, $\rho^2 \in L^\infty(\Omega)$, ν_N is the outward normal to the Neumann boundary, and the diffusion matrix $A(x)$ is symmetric and uniformly positive definite ($c_1|\xi|^2 \le A(x)\xi \cdot \xi \le c_2|\xi|^2$).

This problem can be approximated by different two-dimensional models. The simplest one is the so called *zero-order reduced model*, which is based on the hypothesis that *the exact solution is almost constant with respect to* x_3, which effectively means that we find solution in the subspace

$$V_0^{(0)} := \Big\{ v \in V_0 \mid \exists\, \widehat{v} \in H^1(\widehat{\Omega}) \text{ such that } \widehat{v} = 0 \text{ on } \widehat{\Gamma}_D \text{ and } v(x) = \widehat{v}(\widehat{x}) \text{ for a.e. } x = (\widehat{x}, x_3) \in \Omega \Big\}. \tag{2}$$

This assumption leads to the simplified 2D model

$$\begin{aligned} -\widehat{\nabla}\cdot(d(\widehat{x})\,\widetilde{A}_p(\widehat{x})\widehat{\nabla}\widehat{u}^{(0)}) + d(\widehat{x})\widetilde{\rho}^2(\widehat{x})\,\widehat{u}^{(0)} &= d(\widehat{x})\,\widehat{f}(\widehat{x}) \quad \text{in } \widehat{\Omega}\,, \\ \widehat{u}^{(0)} &= 0 \qquad\qquad\quad \text{on } \widehat{\Gamma}_D, \end{aligned} \tag{3}$$

Theorem 1 ([13], ii). *Assume ρ strictly positive, $B := A^{-1}$ and $b_3 := \{b_{31}\,,\ b_{32}\,,\ b_{33}\}^T$. The error of the zero-order reduced model is subject to the following estimate:*

$$\|\!|\, u - \widehat{u}^{(0)} \,|\!\| \le M_{mod} := \sqrt{M_1^2 + M_2^2} + M_3, \tag{4}$$

where

$$M_1 := \Bigg(\int_{\widehat{\Omega}} d\,(\widehat{x})\,(\widetilde{B}_{p,0}\widetilde{A}_{p,0} - I)\widehat{\nabla}\widehat{u}^{(0)}\cdot\widetilde{A}_{p,0}\widehat{\nabla}\widehat{u}^{(0)}\,d\widehat{x} + \int_{\Omega}\Big(b_{33}\psi(x)^2 + 2\Big(b_3\cdot(\widetilde{A}_{p,0}\widehat{\nabla}\widehat{u}^{(0)},0)\Big)\psi(x)\Big)\,dx\Bigg)^{1/2},$$

$$M_2^2 := \Big\|\rho^{-1}\Big((f-\widehat{f}_0) - (\rho^2-\widetilde{\rho}_0^2)\,\widehat{u}^{(0)} - \frac{\widehat{\nabla}d(\widehat{x})}{d(\widehat{x})}\cdot\widetilde{A}_{p,0}\widehat{\nabla}\widehat{u}^{(0)} + \frac{\partial\psi}{\partial x_3}\Big)\Big\|^2_{L_2(\Omega)}.$$

and

$$M_3 := c_1^{-1}\,C_{\Gamma_N}\,\|F - (\widetilde{A}_{p,0}\widehat{\nabla}\widehat{u}^{(0)},0)\cdot\nu_N - \psi\nu_{N,3}\|_{L_2(\Gamma_N)}.$$

Here C_{Γ_N} denotes the trace constant and ψ is an auxiliary function

We outline that the right hand side of (4) contains only components of 2D solution and the function $\psi \in L_2(\Omega)$, $\psi \in L_2(\Gamma_N)$, $\frac{\partial\psi}{\partial x_3} \in L_2(\Omega)$, which can be taken arbitrary. Any ψ yields an upper bound, but getting the sharpest bound requires minimization over ψ.

The theorem generalizes earlier results [18] devoted to analysis of modeling errors of dimensionally reduced linear elliptic equations. We note that the estimates obtained in publications [13, 18] are also applicable to thin structures with non-planar faces. Plates with plane parallel faces form a particular class of such problems. For them, the modeling error estimate converts to

$$M_{mod} \le \sqrt{\frac{d_0}{3}}\left(\int_{\widehat{\Omega}}\frac{F_\oplus^2 + F_\ominus^2 - F_\oplus F_\ominus}{a_{33}}\,d\widehat{x}\right)^{1/2} + C_\Omega\|f-\widetilde{f}\|_{L^2(\Omega)}\,,$$

where d_0 is the thickness parameter and $F_\oplus(\widehat{x})$ and $F_\ominus(\widehat{x})$ are the forces acting on the upper and lower faces and $\widetilde{f}$ is f averaged with respect to x_3–coordinate. If, in addition, we assume that $f=0$, $a_{33}=1$ and $F_\oplus = F_\ominus = F(\widehat{x})$, then $M_{mod} \le \sqrt{\frac{d_0}{3}}\,\|\widehat{F}\|_{L^2(\widehat{\Omega})}$, which is exactly the estimate obtained in [1, 2] for

zero–order models. In order to derive a more sophisticated reduced problem, models of higher order should be used. The corresponding modeling error estimates are presented in [13].

2.2 Linear Elasticity

In the linear elasticity theory, we consider the following elliptic system for the displacement u and stress σ:

$$\sigma(x) = \mathbb{L}\varepsilon(u)(x), \quad \varepsilon(u)(x) = \tfrac{1}{2}(\nabla u(x) + (\nabla u(x))^T) \text{ in } \Omega, \tag{5}$$

$$\nabla \cdot \sigma(x) + f(x) = 0, \qquad f = (f_1(\widehat{x}),\, f_2(\widehat{x}),\, 0) \quad \text{in } \Omega, \tag{6}$$

$$\sigma(x)\nu = F(x), \qquad F = (F_1(\widehat{x}),\, F_2(\widehat{x}),\, 0) \qquad \text{on } \Gamma_2, \tag{7}$$

$$\sigma\nu = 0 \quad \text{on } \Gamma_N, \qquad u(x) = u_0(x) \qquad \text{on } \Gamma_1. \tag{8}$$

Let $\widehat{u} = (\widehat{u}_1, \widehat{u}_2)$ and $\widehat{\sigma}$ denote the solutions of a simplified 2D problem based on "a priori plane stress assumptions":

$$\sigma_{13} = \sigma_{23} = \sigma_{33} = 0, \quad \sigma_{\alpha\beta} = \sigma_{\alpha\beta}(\widehat{x}), \ \alpha, \beta = 1, 2, \quad u_\alpha = u_\alpha(\widehat{x}).$$

Thus, it is required to find $\widehat{u} = (\widehat{u}_1(\widehat{x}), \widehat{u}_2(\widehat{x}))$ and $\widehat{\sigma} = \widehat{\sigma}_{\alpha\beta}(\widehat{x}) \quad \alpha, \beta = 1, 2$ that satisfy a (simplified system) *plane stress problem*

$$\widehat{\sigma} = \widehat{\mathbb{L}}\,\widehat{\varepsilon}, \qquad \widehat{\varepsilon} = \tfrac{1}{2}(\widehat{\nabla}\widehat{u} + (\widehat{\nabla}\widehat{u})^T) \qquad \text{in } \widehat{\Omega}, \tag{9}$$

$$\widehat{\nabla} \cdot \widehat{\sigma} + \widehat{f} = 0, \quad \widehat{\nabla} = (\tfrac{\partial}{\partial x_1}, \tfrac{\partial}{\partial x_2}), \quad \widehat{f} = (f_1(\widehat{x}), f_2(\widehat{x})) \quad \text{in } \widehat{\Omega}, \tag{10}$$

$$\widehat{u} = \widehat{u}_0, \quad \widehat{u}_0 = (u_{01}(\widehat{x}), u_{02}(\widehat{x})) \qquad \text{on } \widehat{\Gamma}_1, \tag{11}$$

$$\widehat{\sigma}\widehat{\nu} = \widehat{F}, \quad \widehat{F} = (F_1(\widehat{x}), F_2(\widehat{x})) \qquad \text{on } \widehat{\Gamma}_2. \tag{12}$$

If the media is isotropic, then

$$\mathbb{L}\,\varepsilon = K_0\, \mathrm{tr}(\varepsilon)\,\mathbb{I} + 2\mu\boldsymbol{\varepsilon}^D, \quad \widehat{\sigma} = \widehat{\mathbb{L}}\,\widehat{\varepsilon} = \widehat{K}_0\,\widehat{\mathrm{tr}}(\widehat{\varepsilon})\,\widehat{\mathbb{I}} + 2\mu\varepsilon^D, \tag{13}$$

where K_0 and μ are positive (elasticity) constants, tr is the first invariant of a tensor, $\boldsymbol{\varepsilon}^D$ is the deviator of ε, $\mathbb{I}$ is the unit tensor, and $\widehat{K}_0 = \frac{9K_0\mu}{3K_0+4\mu}$.

Now, we reconstruct an approximate 3D solution as follows:

$$\widetilde{u} = (\widehat{u}_1, \widehat{u}_2, \phi(x_1, x_2, x_3)), \ \ \widetilde{\sigma}_{\alpha\beta} = \widehat{\sigma}_{\alpha\beta}, \ \ \widetilde{\sigma}_{3\alpha} = 0,$$

where $\phi \in H^1(\Omega)$ and satisfies the same boundary conditions as u_{03} on the Dirichlet part of Γ (the function ϕ is in our disposal). The error encompassed in this dimensionally reduced model is estimated as follows:

Theorem 2 ([11]). *Under the assumptions made above*

$$C_\varepsilon \,\|\, \varepsilon(\widetilde{u} - u)\|_\Omega^2 + C_\tau \,\|\, \widetilde{\sigma} - \sigma\|_\Omega^2 \le$$
$$\le \left(\frac{K_0}{2} + \frac{2\mu}{3}\right) \int_\Omega \left(\rho(\widehat{u}_{1,1} + \widehat{u}_{2,2}) + \phi_{,3}\right)^2 dx + \mu \int_\Omega \left(\phi_{,1}^2 + \phi_{,2}^2\right) dx \quad (14)$$

where $\rho = \frac{3K_0 - 2\mu}{3K_0 + 4\mu} = \frac{\lambda}{\lambda + 2\mu} = \frac{\nu}{1-\nu}$, $C_\varepsilon = \min\{2\mu, 3K_0\}$, $C_\tau = \frac{1}{\max\{2\mu, 3K_0\}}$.

We see that the right hand side contains only $\widehat{u}$ (the solution of (9)–(13)) and the function ψ, which can be considered as a reconstruction of the third velocity component.

The estimate (14) shows that minimum of the right hand side may be attained not for $\psi = 0$. Finding the best ψ requires minimization of the estimate what amount solving a diffusion type problem. Any approximate solution of the latter problem gives a guaranteed error bound.

Recently, computable estimates of modeling errors has been derived for Kirchhoff-Love plates (see [17]). These estimates does not contain 3D constants and preserve asymptotic properties under the same assumptions that are used in convergence analysis.

3 Errors of Defeatured Models

Defeaturing is a way to simplify models by neglecting some details, which seem to be insignificant. In this case, the modeling error is $u_\epsilon - u$, where u_e is the solution of a simplified model and ϵ is the defeaturing parameter such that $u_0 = u$. For the problem (1), the corresponding error estimates has been derived in [14, 16]. Since approximation errors can be estimated with the help of the same mathematical method as modeling errors, we obtain the following estimate:

$$\|\nabla(u - u_{\varepsilon,h})\|_A \le E_{disc}^{\varepsilon,h} + E_{mod}^{\varepsilon}, \quad (15)$$

which includes both approximation error $E_{disc}^{\varepsilon,h}$ (associated with the mesh $\mathscr{T}_h$ used to solve a simplified model) and modeling error E_{mod}^{ε} (generated by a simplified matrix A_ϵ) defined by the relations

$$E_{disc}^{\varepsilon,h} := \|\nabla(u_\varepsilon - u_{\varepsilon,h})\|_A \le \kappa_1 \, \mathscr{M}_\Omega(u_{\varepsilon,h}, y, \beta), \quad (16)$$

$$E_{mod}^{\varepsilon} := \|\nabla(u - u_\varepsilon)\|_A \le \kappa_\varepsilon \left(\frac{\kappa_2}{2} \mathscr{M}_\Omega^2(u_{\varepsilon,h}, y, \beta) + \int_\Omega f \, u_{\varepsilon,h}\right)^{1/2}, \quad (17)$$

where $\mathscr{M}_\Omega$ is the functional error majorant for the simplified boundary value problem. It is defined by the following relation:

$$\mathscr{M}^2_{\Omega}(u_{\varepsilon,h}, y, \beta) := (1+\beta)\|A_\varepsilon \nabla u_{\varepsilon,h} - y\|^2_{A^{-1}_\varepsilon} + \frac{1+\beta}{\beta} C^2_{\Omega} \|\nabla \cdot y + f\|^2_{\Omega}.$$

Here, $y \in H(\Omega, \mathrm{div})$ is an arbitrary vector-valued function, β is an arbitrary positive number, $C_\Omega := c_1^{-2} C^2_{F\Omega}$, where $C_{F\Omega}$ is the Friedrichs constant for the domain Ω, $\kappa_1^2 := 1 + \varrho(\Lambda_\varepsilon - I)$, I is the identity matrix, $\Lambda_\varepsilon := A_\varepsilon^{-1/2} A A_\varepsilon^{-1/2}$, $\kappa_\varepsilon^2 := \frac{2\kappa_2}{2\kappa_2 - 1}\, \varrho(\Lambda_\varepsilon + \Lambda_\varepsilon^{-1} - 2I)$, ϱ denotes the maximal eigenvalue, and κ_2 is easy computable (see [14], (18)).

Explicit knowledge about the overall error and values of the two parts of it generates combined "modeling-discretization" adaptive numerical strategy. For defeatured models high efficiency of this strategy has been numerically confirmed in [14]. We note that similar methods are quite natural for computations with dimensionally reduced or homogenized models.

For the class of the dimension reduction, a similar combined modeling-discretization estimate of the total error is presented by

Theorem 3 ([19]). *The total error is bounded from above by the sum*

$$|||\, \nabla(u - \widehat{u}^{(m)}_h)\, ||| \le E^{(m),h}_{mod} + E^h_{app}, \tag{18}$$

where m is the order of the 2D-model, $E^{(m),h}_{mod}$ and E^h_{app} represent the modeling and the discretization parts of the error, respectively, and are defined and estimated as follows:

$$E^h_{app} := |||\, \nabla(\widehat{u}^{(m)} - \widehat{u}^{(m)}_h)\, ||| , \tag{19}$$

$$E^{(m),h}_{mod} := |||\, \nabla(u - \widehat{u}^{(m)})\, ||| \le M_{mod}(\widehat{u}^{(m)}_h, y) + E^h_{app}, \tag{20}$$

where for all $y \in H(\Omega, \mathrm{div})$

$$M_{mod}(\widehat{u}^{(m)}_h, y) := \sqrt{\breve{M}^2_1(\widehat{u}^{(m)}_h, y) + \breve{M}^2_2(\widehat{u}^{(m)}_h, y)} + c_1^{-1} C_{\Gamma_N} \breve{M}_3(\widehat{u}^{(m)}_h, y)$$

with

$$\breve{M}_1(\widehat{u}^{(m)}_h, y) := \left(\int_\Omega (\nabla \widehat{u}^{(m)}_h - A^{-1} y) \cdot (A \nabla \widehat{u}^{(m)}_h - y)\, dx \right)^{1/2},$$

$$\breve{M}_2(\widehat{u}^{(m)}_h, y) := \left\| \rho^{-1} \left(\nabla \cdot y - \rho^2 \widehat{u}^{(m)}_h + f \right) \right\|_{L_2(\Omega)},$$

and

$$\breve{M}_3(\widehat{u}^{(m)}_h, y) := \|F_N - y \cdot \nu_N\|_{L_2(\Gamma_N)}.$$

The majorant of the approximation error E^h_{app} depends on the reduced model. In the case of the zero-order model, it holds

$$E^h_{app} \le \widehat{M}_{app}(\widehat{u}^{(0)}_h, \hat{y}) := \widehat{M}_1(\widehat{u}^{(0)}_h, \hat{y}) + C_{F,\widehat{\Omega}}\, c_1^{-1}\, \widehat{M}_2(\widehat{u}^{(0)}_h, \hat{y})$$

with

$$\widehat{M}_1(\widehat{u}_h^{(0)}, \hat{y}) := \left(\int_{\widehat{\Omega}} \left(\widehat{\nabla} \widehat{u}_h^{(0)} - \left(d(\widehat{x})\, \widetilde{A}_p \right)^{-1} \hat{y} \right) \cdot \left(d(\widehat{x})\, \widetilde{A}_p \widehat{\nabla} \widehat{u}_h^{(0)} - \hat{y} \right) dx \right)^{1/2}$$

and $$\widehat{M}_2(\widehat{u}_h^{(0)}, \hat{y}) := \| \widehat{\nabla} \cdot \hat{y} - d(\widehat{x})\, \widetilde{\rho}^2\, \widehat{u}_h^{(0)} + \widehat{f} \|_{L_2(\widehat{\Omega})}$$

for all $\hat{y} \in H(\widehat{\Omega}, \mathrm{div})$.

4 Errors of Homogenized Models

We consider a bounded domain $\Omega \subseteq \mathbb{R}^d$ such that $\Omega = \sum_i \Pi_i^\varepsilon$, where $i = (i_1, i_2, \ldots i_d)$ is the multiindex, $\Pi_i^\varepsilon = x_i + \widehat{\Pi}$ is a "translated cell", and x_i is the global coordinate of its reference point. Any cell is generated by the etalon cell $\widehat{\Pi}$. In $\widehat{\Pi}$ (which defines the periodic structure), we use local coordinates y. For any Π_i^ε local and global coordinates are joined by the relation $y = \frac{x - x_i}{\varepsilon} \in \widehat{\Pi}$. Let $\widehat{\Pi} = \left( ]0, 1[\right)^d$ be the unit cube. On $\widehat{\Pi}$, we define a matrix function $\widehat{A}(y)$ and assume that $\widehat{A} \in L^\infty(\widehat{\Pi}, M_{sym}^{d \times d})$ and uniformly positive definite. By $A_\varepsilon(x)$, we define the periodical structure on Ω, namely $A_\varepsilon(x) = \widehat{A}\left(\frac{x - x_i}{\varepsilon} \right)$, $x \in \Pi_i^\varepsilon$, where ε is a small parameter (geometrical size of a cell) and consider in Ω the problem $-\nabla \cdot (A_\varepsilon(x) \nabla u_\varepsilon) = f$ with the uniform Dirichlet boundary conditions. It is well known from the homogenization theory (see, e.g., [5–7]) that using the so called "homogenized" matrix A_0, we can define $u_0 \in H_0^1(\Omega)$

$$\int_\Omega A_0 \nabla u_0 \cdot \nabla w dx = \int_\Omega f w dx, \ \forall w \in H_0^1(\Omega) \tag{21}$$

and prove that $u_\varepsilon \rightharpoonup u_0$ in $H_0^1(\Omega)$ (which means that solution of the homogenized model can be viewed as an approximation of u_ϵ). Moreover, it has been proved that it is possible to correct u_0 and construct a function w_ε (a corrected approximation of u_ε) such that $\| u_\varepsilon - w_\varepsilon \|_{H^1(\Omega)} \leq c \sqrt{\varepsilon}$. A computable bound of the modeling error $w_\epsilon - u_\epsilon$ can be derived by the same methods that has been applied to dimension reduction models. It is expressed throughout two quantities:

$$M_1^2 := \varepsilon^{2s} |i| \, \| \eta(y) \|^2_{\widehat{A}^{-1} \widehat{\Pi}} + 2 \varepsilon^s \int_{\widehat{\Pi}} M \cdot \eta(y)\, dy$$
$$+ \varepsilon^s \Big(\sum_{j=1}^{2} \frac{1}{\lambda_j} \Big\| \eta^j(y) - \frac{1}{|\widehat{\Pi}|} \int_{\widehat{\Pi}} \eta^j \, dy \Big\|^2_{\widehat{\Pi}} + \lambda \cdot \mu \Big)$$

$$+\sum_i \int_{\Pi_i^\varepsilon} \widehat{A}\left(\frac{x-x_i}{\varepsilon}\right) g(x)\cdot g(x)\, dx, \quad \text{and} \quad M_2^2 = \varepsilon^{2s}\, \widetilde{C}\, \| \nabla\cdot_y \eta \|^2_{\widehat{\Pi}}.$$

Here $s \geq 1$ $\lambda = (\lambda_1, \dots \lambda_d)$ is an arbitrary vector, $\eta(y) \in H(\widehat{\Pi}, \text{div})$ is an arbitrary function, M is the vector of mean values on the cells, i.e., $M := \sum_{\mathbf{i}} \langle g \rangle_{\Pi_{\mathbf{i}}^\varepsilon}$, $\mu^{1/2} := \sum_{\mathbf{i}} \| g - \langle g \rangle_{\Pi_{\mathbf{i}}^\varepsilon} \|_{\Pi_{\mathbf{i}}^\varepsilon}$, and $g(x)$ is a known function defined by ε, A_ε, A_0, $\nabla u_0(x)$, and other parameters of the model (see details in [15]).

Theorem 4 ([15]). *Error of the homogenized model is subject to the estimate*

$$\| \nabla(u_\varepsilon - w_\varepsilon) \|^2_{A_\varepsilon} \leq (1+\beta)\, M_1^2 + \frac{1+\beta}{\beta} M_2^2.$$

We note that the parameters λ, β, and the function η (which is defined on the cell $\widehat{\Pi}$) can be used in order to minimize the right hand side of the estimate.

References

1. I. Babuška, I. Lee, and C. Schwab. On the a posteriori estimation of the modeling error for the heat conduction in aplate and its use for adaptive hierarhical modeling. *Appl. Numer. Math.*, 14(1–3):5–21, 1994.
2. I. Babuška and C. Schwab. A posteriori error estimation for hierarchic models of elliptic boundary value problems on thin domains. *SIAM J. Numer. Anal.*, 33:221–246, 1996.
3. I. Babuška and M. Vogelius. On a dimensional reduction method I. The optimal selection of basis functions. *Math. Comp.*, 37:31–46, 1981.
4. I. Babuška and M. Vogelius. On a dimensional reduction method III. A posteriori error estimation and adaptive approach. *Math. Comp.*, 37:361–384, 1981.
5. A. Bensoussan, J.-L. Lions, and G. Papanicolaou. *Asymptotic Analysis for periodic structures*. Amsterdam: North-Holland, 1978.
6. M. Chipot. *Elliptic Equations: An Introductory Course*. Birkhauser Verlag AG, 2009.
7. V. Jikov, S. Kozlov, and O. Oleinik. *Homogenization of differential operators and integral functionals*. Springer, Berlin, 1994.
8. O. Mali and S. Repin. Estimates of accuracy limit for elliptic boundary value problems with uncertain data. *Adv.Math. Sci. Appl.*, 19(2):525–537, 2009.
9. P. Neittaanmäki and S. Repin. *Reliable methods for computer simulation*, volume 33 of *Studies in Mathematics and its Applications*. Elsevier Science B.V., Amsterdam, 2004. Error control and a posteriori estimates.
10. S. Repin. A posteriori error estimation for variational problems with uniformly convex functionals. *Math. Comp.*, 69:481–600, 2000.
11. S. Repin. Estimates for errors in two-dimensional models of elasticity theory. *J. Math. Sci. (NY)*, 106(3):27–41, 2001.
12. S. Repin. *A posteriori estimates for partial differential equations*. Walter de Gryter, Berlin, 2008.
13. S. Repin and T. Samrowski. Estimates of dimension reduction errors for stationary reaction-diffusion problems. *J. Math. Sci. (NY)*, 173(6):803–821, 2011.
14. S. Repin, T. Samrowski, and S. Sauter. Combined a posteriori modeling-discretization error estimate for elliptic problems with complicated interfacees. *M2AN, accepted for publication*.

15. S. Repin, T. Samrowski, and S. Sauter. Two-sided estimates of the modeling error for a homogenization problem. Preprint 12–2012, University Zurich.
16. S. Repin, T. Samrowski, and S. Sauter. A posteriori estimates fir modelling errors of the stationary diffusion equation with complicated coefficients. *PAMM*, 10(1):571–572, 2010.
17. S. Repin and S. Sauter. Computable estimates of the modeling error related to Kirchhoff–Love plate model. *Anal. Appl.*, 8(4).
18. S. Repin, S. Sauter, and A. Smolianski. A posteriori estimation of dimension reduction errors for elliptic problems in thin domains. *SIAM J. Numer. Anal.*, 42:1435–1451, 2004.
19. T. Samrowski. Combined error estimates in the case of the dimension reduction. Preprint 16–2011, University Zurich.
20. C. Schwab. A-posteriori modeling error estimation for hierarchic plate models. *Numer. Math.*, 74(2):221–259, 1996.

Anisotropic Finite Elements for Fluid-Structure Interactions

T. Richter

Abstract In this work, we present an adaptive finite element method for the numerical simulation of fluid-structure interaction problems using anisotropic meshes. By formulating the coupled problem in a fully monolithic variational Arbitrary Lagrangian Eulerian framework, sensitivities for guiding goal-oriented error estimation are easily at hand. The errors are locally estimated separately in the different element-coordinate directions. This allows for a directional splitting of elements and the generation of anisotropic meshes. The goal-oriented error estimator is applied to a stationary benchmark problem coupling the incompressible Navier-Stokes equations with a nonlinear hyper-elastic material law.

1 Introduction

Fluid-Structure interaction (FSI) is part of various technical problems. Most of these application problems (e.g. in aerodynamics or hemodynamics) are three dimensional. The cost connected with the simulation of three dimensional FSI-problems is immense. The use of locally refined meshes is an effective remedy in the case of large three dimensional problems. While adaptive finite element methods have a long tradition and are well established for flow [18] and structural [1, 25] problems, the consideration of multi-physics problems is a recent development.

In this work we develop a goal-oriented error estimator for FSI-problems which allows for a separation of the anisotropic error influences. Besides being used as a estimator, we construct optimal anisotropic meshes. This is accomplished by introducing anisotropic mesh refinements, where an element can be split in particular directions only.

T. Richter (✉)
Institute of Applied Mathematics, University of Heidelberg, INF 294, 69120 Heidelberg, Germany
e-mail: thomas.richter@iwr.uni-heidelberg.de

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_7,

First work on a posteriori error estimation with help of a duality technique has been done by Eriksson, Johnson and co-workers, see [9] for a survey. Becker and Rannacher [5] further developed this approach into a computation-based method, the *Dual Weighted Residual*-method (DWR), where the sensitivities are approximated as discretized adjoint equations. Further works on goal-oriented adaptivity and duality techniques for partial differential equations are found in [12, 19].

Extending the concept of error estimation to anisotropic meshes requires the estimation of optimal local directions which align the finite element mesh with dominant anisotropies. One separates between algorithms which construct new meshes with the help of mesh generators and algorithms which are based on a directional refinement of a given mesh by locally splitting elements in the different anisotropic directions. We follow this second way, which is more easy since we do not need identify an optimal mesh metric [8] but only need to analyze the error in the two or three (in 3d) different directions.

Usually anisotropic error estimation is based on an analysis of the solution's Hessian. However this approach is closely linked to interpolation error estimates and estimates in the energy norm and not directly suitable for sensitivity based error estimations. Several authors [10, 15, 17, 24] merge Hessian based anisotropy detection with goal-oriented error estimation in terms of local optimization problems. In this work, we present a uniform approach which directly estimates the directional errors using the DWR-method [20].

Key to goal oriented error estimation are the sensitivities with regard to the considered error goals. Due to the free surface character, these adjoint information are difficult to obtain for coupled fluid-structure interaction problems. By introducing the *Arbitrary-Lagrangian-Eulerian* coordinates (ALE) to map the flow problem onto a fixed reference domain [3, 11, 14, 16], the coupled system can be written in a closed variational formulation and is at hand to sensitivity analysis.

In the remainder of this paper we first present a consistent variational formulation for the coupled fluid-structure interaction problem (Sect. 2) developed in [22] and shortly describe the finite element discretization. Then, in Sect. 3 we present the anisotropic error estimator used for estimation and we describe the anisotropic mesh adaptation scheme. Finally, in Sect. 4 we analyze a numerical benchmark to highlight the scope of the presented technique.

2 Governing Equations and the Finite Element Discretization

By $\Omega \subset \mathbb{R}^d$ with $d = 2, 3$ we denote a domain. This domain is split into a fluid-part Ω_f and into a solid part Ω_s, each domains in $\mathbb{R}^d$. It holds $\Omega_f \cap \Omega_s = \emptyset$ and $\bar{\Omega} = \bar{\Omega}_f \cup \bar{\Omega}_s$. By $\Gamma_i := \bar{\Omega}_f \cap \bar{\Omega}_s$ we denote the *interface*. In Ω_f the incompressible Navier-Stokes equations are given for the fluid's velocity $v_f : \Omega_f \to \mathbb{R}^d$ and pressure $p_f : \Omega_f \to \mathbb{R}$. In Ω_s an elastic material law is given to describe the solid's deformation $u_s : \Omega_s \to \mathbb{R}^d$. Figure 1 shows a typical configuration of

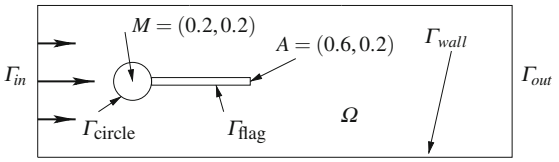


Fig. 1 Typical configuration for a fsi-problem: flow around rigid circle Γ_{circle} with an elastic beam Γ_{flag}

a fsi-problem, where the flow encloses an elastic obstacle. The big challenge of fluid-structure interaction is the deformation of the domains Ω_f and Ω_s under load: the fluid's forces on the obstacle will cause a deformation u_s of the solid $\Omega_s \to \Omega_s^l$. Consequently, the flow domain will move along $\Omega_f \to \Omega_f^l$. The layout of this new configuration is not known, instead it must be considered as unknown part of the solution.

In the following we use the standard notation for the Lebesgue space $L^2(\Omega)$ and the Sobolew space $H^1(\Omega)$ of functions with square integrable weak derivatives. With $H_0^1(\Omega)$ we denote those H^1-functions with trace zero on $\partial\Omega$. By $\|\cdot\|_\Omega$ and $(\cdot,\cdot)_\Omega$ we denote the L^2-norm and inner product in Ω.

2.1 Governing Equations

Here, we shortly present the variational formulation used to model the coupled fluid-structure interaction problem. Details on the derivation of the coupled system are found in [22]. We assume that the flow domain Ω_f is governed by the incompressible Navier-Stokes equations and the solid domain Ω_s by a hyperplastic compressible material of St. Venant Kirchhoff [13] type. To cope with the dilemma of moving and not-matching subdomains when coupling the—usually Eulerian—flow problem with the—usually Lagrangian—structure model, we formulate the coupled system in *Arbitrary Lagrangian Eulerian* (ALE) coordinates by mapping the flow problem onto an artificial coordinate system which is aligned with the solid problems. These coordinates have first been introduced to model free-surface flow problems [14] but then been widely adapted to derive monolithic formulations for fluid-structure interactions, see [3, 7, 11, 26] among many others. Details on the variational formulation used throughout this work are found in [22]:

Problem 1 (FSI-problem in ALE-coordinates). *Find*

$$U := \{v, u, p\} \in X := H_0^1(\Omega_f) \times H_0^1(\Omega_f \cup \Omega_s) \times L^2(\Omega_f) \setminus \mathbb{R},$$

such that

$$A(U)(\Phi) = 0 \quad \forall \Phi := \{\phi, \psi, \xi\} \in X, \tag{1}$$

with

$$A(U)(\Phi) := (J\hat{\sigma}_f F^{-T}, \nabla\phi)_{\Omega_f} + (J\rho F^{-1} v \cdot \nabla v, \phi)_{\Omega_f} + (\operatorname{div}(JF^{-1}v), \xi)_{\Omega_f} + (F_s \Sigma_s, \nabla\phi)_{\Omega_s} + (\nabla u, \nabla\psi)_{\Omega_f},$$

where by $F := \nabla u$ we define the deformation gradient, by J its determinant. With $\hat{\sigma}_f$ and Σ_s we denote the fluid's and solid's stress tensor. See [22] for details.

2.2 Finite Element Discretization

Discretization of Problem 1 will be by means of equal order finite elements for all variables: pressure, velocity and (solid's and fluid's) deformation. First, let Ω_h be a triangulation of the domain Ω into open quadrilaterals or hexahedrals which—in addition to the usual shape-regularity assumption—is build in such a way, that the fluid-structure interface Γ_i is not cut by any element $K \in \Omega_h$. Then, on Ω_h we define by V_h the space of continuous functions on Ω which are piece-wise polynomial of degree $r \in \mathbb{N}$ on every element $K \in \Omega_h$. To allow for general element (not necessarily rectangular) and higher order approximation of curved boundaries, these spaces are assembled by iso-parametric transformation from a given reference element. Local mesh refinement is realized by introducing hanging nodes: degrees of freedom living on edges or faces, where not all adjacent elements are refined are replaced by interpolations from the neighboring degrees of freedom, see [20]. As equal-order finite elements for all solution components are utilized, we need to stabilize the inf-sup condition. Here, we use the *local projection stabilization* (LPS) [4] having an easy diagonal structure as compared to the traditional PSPG/SUPG stabilization techniques. With $A_h(\cdot)(\cdot) := A(\cdot)(\cdot) + S_{\text{lps}}(\cdot)(\cdot)$ where by $S_{\text{lps}}(\cdot)(\cdot)$ we denote the LPS-terms (see [22]) the discrete solution $U_h \in X_h$ is given as

$$A_h(U_h)(\Phi_h) = F(\Phi_h) \quad \forall \Phi_h \in X_h. \tag{2}$$

Problem (2) is a highly nonlinear complex system of partial differential equations. We iterately solve these equations by a Newton method. The arising linear systems are then solved by help of a multigrid-preconditioned GMRES-iteration. See [21] for a detailed description on the solution methods.

3 Anisotropic Error Estimation

In order to allow for anisotropic finite elements with high aspect ratio we relax the shape-regularity condition [20]. Anisotropic meshes are realized by hierarchical refinement. In Fig. 2 we show possible (anisotropic) refinement types. A good

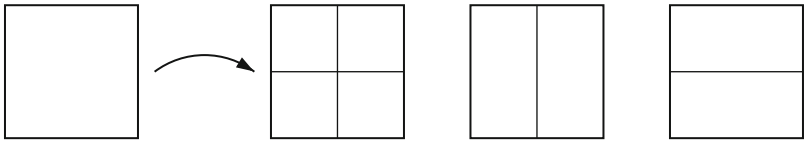


Fig. 2 Possibilities of mesh refinement (in 2d): isotropic refinement into four new elements, anisotropic refinement into two element by splitting in x-direction or y-direction

overview of anisotropic finite element analysis is given in [2]. A mesh with local anisotropic refinement is shown in Fig. 4.

3.1 Error Estimation and Anisotropy Detection

The anisotropic error estimator closely follows the works [20, 22]. While in [20] the basic concept of anisotropic splitting is introduces, we derive in [22] the fundamental variational formulation for the fluid-structure interaction problem and discuss the adjoint equations and adjoint interface conditions in the coupled problem. The main result is stated in the following theorem:

Theorem 1 (Anisotropic error estimation). *Let $U \in X$ and $U_h \in X_h$ be solutions of (1) and (2) and $Z_h \in X_h$ be the linearized adjoint solution with respect to a given error functional $J : X \to \mathbb{R}$:*

$$A_h'(U_h)(\Phi, Z_h) = J'(U_h)(\Phi_h) \quad \forall \Phi_h \in X_h. \tag{3}$$

Then, the goal-error $J(U) - J(U_h)$ is estimated as:

$$J(U) - J(U_h) = \eta_h^x(U_h, Z_h) + \eta_h^y(U_h, Z_h) + \eta_h^z(U_h, Z_h) + \mathscr{R}_{split} \tag{4}$$

where the directional errors are estimated as

$$\eta_h^i = \frac{1}{2}\left\{J'(U_h)(\pi_h^d U) - A'(U_h)(\pi_h^d U_h, Z_h)\right\} - \frac{1}{2}A(U_h)(\pi_h^d Z_h) + \mathscr{R}^d, \quad d = x, y, z.$$

By $\mathscr{R}_{split}$ and $\mathscr{R}^d$ we denote remainders of third order in the error and by π_h^d discrete fluctuation operators in direction d, given by $\pi_h^d := \mathrm{id} - \mathrm{I}_*^d$, where by $\mathrm{I}_*^d : V_h \to V_*^d$ we denote the discrete interpolation into a space V_*^d which has higher approximation order in direction d.

A proof of this fundamental theorem is found in [5], the extension to anisotropic splitting in [20, 22]. The fundamental principle of splitting the discretization error into the different mesh directions is by means of partially discretized spaces V_h^{d} which are used to separate the directional influences. Estimates with regard to these spaces are obtained by local and discrete fluctuation operators based on

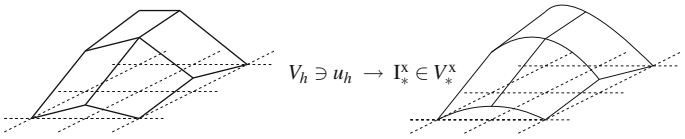


Fig. 3 Discrete interpolation operator into space of higher polynomial degree in x-direction V_*^{x} used to separate anisotropic influences

reconstruction of the solution using a higher interpolation order. See Fig. 3 for a sketch of this construction and [20] for details on the analysis.

3.2 Adaptive Algorithm

As result of the estimator (4) we immediately get a local splitting of the error into different directions. Mesh adaptation and deciding about the optimal direction of element-splitting is done in one step. Given an initial mesh $\Omega_h^{(1)}$ we iterate: for $i \geq 1$

1. Solve $U_h^{(i)}, Z_h^{(i)} \in X_h^{(i)}$.
2. Estimate the error: $\eta_h^{(i)} := \eta_h^{\mathrm{x},(i)}(U_h^{(i)}, Z_h^{(i)}) + \eta_h^{\mathrm{y},(i)}(U_h^{(i)}, Z_h^{(i)}) + \eta_h^{\mathrm{z},(i)}(U_h^{(i)}, Z_h^{(i)})$
3. Stop, if $\eta_h^{(i)} < \mathrm{tol}$.
4. Refine mesh-element $K_j \in \Omega_h^{(i)}$ in d-direction, if $\eta_{K_j}^{\mathrm{d},(i)} > \alpha \bar{\eta}_h^{(i)}$.

By $\eta_K^{\mathrm{d},(i)}$ we denote a localization of the error estimate $\eta_h^{\mathrm{d},(i)}$ onto the element $K \in \Omega_h$. By $\bar{\eta}_h^{(i)} := \eta_h^{(i)}/|\Omega_h^{(i)}|$ we denote the average error. Elements are refined, if the local error is above the average with a certain threshold $\alpha > 0$. This algorithm checks for refinement separately in every direction. If the directional error is above average in multiple directions we also split the element in multiple directions up to isotropic refinement. See [20] for details.

4 Numerical Results

In this section, we study the FSI-I benchmark problem as introduced and extensively analyzed in the collection [7]. The flow around a cylinder (diameter $D = 0.1\,\mathrm{m}$), with an attached elastic beam (length $L = 0.35\,\mathrm{m}$) is simulated. Figure 1 shows a sketch of the configuration. The problem is driven by a parabolic inflow profile with mean velocity $\bar{v}_f = 0.2\,\mathrm{m/s}$. Viscosity of the fluid is set to $\nu = 10^{-3}\,\mathrm{m}^2/\mathrm{s}$, fluid's and solid's density are both chosen as $\rho_f = \rho_s = 10^3\,\mathrm{kg/m}^3$. Finally, the solid is governed by a St. Venant-Kirchhoff material with Poisson ratio $\nu_s = 0.4$ and Lamé-coefficient $\mu_s := 2 \cdot 20^6\,\mathrm{kg/(ms}^2)$. As quantity of interest, we measure the drag-coefficient of the obstacle.

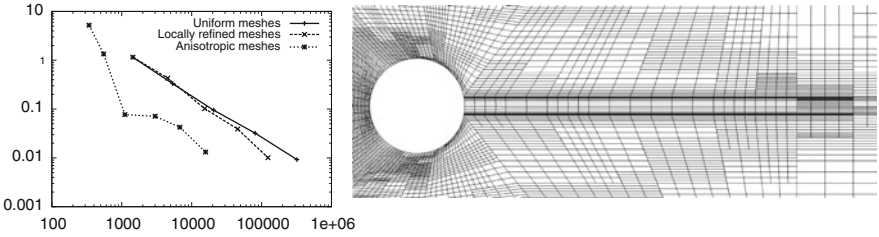


Fig. 4 *Left*: convergence history for the drag-coefficient on uniform meshes, isotropic adaptive meshes and anisotropic adaptive meshes. *Right*: cut-out of the finite element mesh used for calculating the drag-coefficient. Maximum aspect ratio 1:50

By simulations on highly refined uniform and adaptive meshes in [22] as well as by comparing with various contributions in [7] we fix the reference value for the drag-coefficient to $J_{\text{drag}} := 14.2940 \pm 5 \cdot 10^{-4}$

The adjoint solution $Z_h \in X_h$ as given by (3) is approximated with the same approximation space as $U_h \in X_h$. The right hand side is given by $J'_{\text{drag}}(U_h)(\Phi)$ and can be implemented in form of non-conforming Dirichlet data (non-conforming w.r.t. the primal problem). See [6, 22] for details.

In Fig. 4 we compare the convergence history of the drag-coefficient using uniform meshes, adaptive but isotropic meshes with anisotropic adaptive meshes. It is observed, that by using anisotropic finite elements, we increase the accuracy by an additional factor of 10 compared to using isotropic adaptive finite elements. Further, in the right sketch in Fig. 4 we show a cut-out of an anisotropic mesh used in this calculation. In [22] further details are given on this benchmark problem. Here, we also analyze the splitting into primal and adjoint residuals of the error estimator. Further we give more detail on the adjoint solutions.

Using piecewise linear finite elements on uniform meshes, the drag-coefficient should converge with second order given sufficient regularity of the solution. Figure 4 however depicts linear convergence only. This order reduction is due to limited regularity induced by the reentrant edges at the interface Γ_i as seen from the fluid domain. Similar results are observed for pure fluid dynamics benchmark problems [6, 23]. By using locally refined meshes without anisotropy information the efficiency of the discretization cannot significantly enhanced. Only by using anisotropic finite elements, we observe a large gain in accuracy using the same number of unknowns.

References

1. Ainsworth, M., Oden, J.T.: A posteriori error estimation in finite element analysis. Computer Methods in Applied Mechanics and Engineering **142**(1–2), 1–88 (1997)
2. Apel, T.: Anisotropic finite elements: Local estimates and applications. Advances in Numerical Mathematics. Teubner, Stuttgart (1999)
3. Bazilevs, Y., Calo, V., Hughes, T., Zhang, Y.: Isogeometric fluid-structure interaction: theory, algorithms, and computations. Comput Mech **43**, 3–37 (2008)

4. Becker, R., Braack, M.: A two-level stabilization scheme for the Navier-Stokes equations. In: e.a. M. Feistauer (ed.) Numerical Mathematics and Advanced Applications, ENUMATH 2003, pp. 123–130. Springer (2004)
5. Becker, R., Rannacher, R.: An optimal control approach to a posteriori error estimation in finite element methods. In: A. Iserles (ed.) Acta Numerica 2001, vol. 37, pp. 1–225. Cambridge University Press (2001)
6. Braack, M., Richter, T.: Solutions of 3D Navier-Stokes benchmark problems with adaptive finite elements. Computers and Fluids **35**(4), 372–392 (2006)
7. Bungartz, H.J., Schäfer, M. (eds.): Fluid-Structure Interaction II. Modelling, Simulation, Optimisation. Lecture Notes in Computational Science and Engineering. Springer (2010)
8. Castro-Diaz, M., Hecht, F., Mohammadi, B., Pironneau, O.: Anisotropic unstructured mesh adaptation for flow simulations. Int. J. Numer. Math. Fluids. **25**, 475–491 (1997)
9. Eriksson, K., Estep, D., Hansbo, P., Johnson, C.: Introduction to adaptive methods for differential equations. In: A. Iserles (ed.) Acta Numerica 1995, pp. 105–158. Cambridge University Press. (1995)
10. Formaggia, L., Perotto, S., Zunino, P.: An anisotropic a-posteriori error estimate for a convection-diffusion problem. Computing and Visualization in Science **4**, 99–2001 (2001)
11. Ghattas, O., Li, X.: A variational finite element method for stationary nonlinear fluid-solid interaction. Journal of Computational Physics **121**, 347–356 (1995)
12. Giles, M., Süli, E.: Adjoint methods for pdes: a posteriori error analysis and postprocessing by duality. Acta Numerica 2002 pp. 145–236 (2002). A. Iserles, ed.
13. Holzapfel, G.: Nonlinear Solid Mechanics: A Continuum Approach for Engineering. Wiley-Blackwell (2000)
14. Hughes, T., Liu, W., Zimmermann, T.: Lagrangian-eulerian finite element formulations for incompressible viscous flows. Computer Methods in Applied Mechanics and Engineering **29**, 329–349 (1981)
15. L. Formaggia, S.M., Perotto, S.: Anisotropic mesh adaptation in computational fluid dynamics: application to the advection-diffusion-reaction and the stokes problems. Appl. Numer. Math. **51**(4), 511–533 (2004)
16. Le Tallec, P., Mouro, J.: Fluid structure interaction with large structural displacement. Comput. Methods Appl. Mech. Engrg. **190**, 3039–3067 (2001)
17. Leicht, T., Hartmann, R.: Anisotropic mesh refinement for discontinuous galerkin methods in two-dimensional aerodynamic flow simulations. Int. J. Numer. Math. Fluids. (2007). To appear
18. Oden, J., Wu, W., Ainsworth, M.: An a posteriori error estimate for finite element approximations of the navier-stokes equations. Computer Methods in Applied Mechanics and Engineering **111**, 185–202 (1993)
19. Paraschivoiu, M., Patera, A.: Hierarchical duality approach to bounds for the outputs of partial differential equations. Comput. Methods Appl. Mech. Engrg. **158**, 389–407 (1998)
20. Richter, T.: A posteriori error estimation and anisotropy detection with the dual weighted residual method. Int. J. Numer. Math. Fluids. **62**(1), 90–118 (2010)
21. Richter, T.: A monolithic multigrid solver for 3d fluid-structure interaction problems. submitted to Siam J. Scientific Computing (2011)
22. Richter, T.: Goal-oriented error estimation for fluid-structure interaction problems. Computer Methods Applied Mechanics and Engineering 223–224, pp. 38–42 (2012)
23. Schäfer, M., Turek, S.: Benchmark computations of laminar flow around a cylinder. (With support by F. Durst, E. Krause and R. Rannacher). In: E. Hirschel (ed.) Flow Simulation with High-Performance Computers II. DFG priority research program results 1993–1995, no. 52 in Notes Numer. Fluid Mech., pp. 547–566. Vieweg, Wiesbaden (1996)
24. Venditti, D., Darmofal, D.: Anisotropic grid adaptation for functional outputs: application to two-dimensional viscous flows. JCP **187**, 22–46 (2003)
25. Verfürt, R.: A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques. Wiley/Teubner, New York-Stuttgart (1996)
26. Wall, W.: Fluid-structure interaction with stabilized finite elements. Ph.D. thesis, University of Stuttgart (1999). Urn:nbn:de:bsz:93-opus-6234

Adaptive Finite Elements for Semilinear Reaction-Diffusion Systems on Growing Domains

C. Venkataraman, O. Lakkis, and A. Madzvamuse

Abstract We propose an adaptive finite element method to approximate the solutions to reaction-diffusion systems on time-dependent domains and surfaces. We derive a computable error estimator that provides an upper bound for the error in the semidiscrete (space) scheme. We reconcile our theoretical results with benchmark computations.

1 Introduction

Our model problem consists of a system of chemicals that are coupled only through the reaction terms and diffuse independently of each other. Given an integer $m \geq 1$, let $\boldsymbol{u}(\boldsymbol{x},t)$ be an $(m \times 1)$ vector of concentrations of chemical species, with $\boldsymbol{x} \in \Omega_t \subset \mathbb{R}^2$, the spatial variable and $t \in [0,T]$, $T > 0$, the time variable. The model we shall consider is of the following form (see [1] for details of the derivation): find u_i, functions from Ω_t into $\mathbb{R}$, such that for $i = 1,\ldots,m$, u_i satisfies

$$\begin{cases} \partial_t u_i(\boldsymbol{x},t) - D_i \Delta u_i(\boldsymbol{x},t) + \nabla \cdot [\boldsymbol{a} u_i](\boldsymbol{x},t) = f_i\left(\boldsymbol{u}(\boldsymbol{x},t)\right), & \boldsymbol{x} \in \Omega_t, t \in (0,T], \\ [\boldsymbol{v} \cdot \nabla u_i](\boldsymbol{x},t) = 0, & \boldsymbol{x} \in \partial\Omega_t, t > 0, \\ u_i(\boldsymbol{x},0) = u_i^0(\boldsymbol{x}), & \boldsymbol{x} \in \Omega_0, \end{cases} \tag{1}$$

The research of CV has been supported by the EPSRC, Grant EP/G010404. This work (AM) is partly supported by the following grants: EPSRC (EP/H020349/1), the LMS grant (R4P2), and the British Council through its UK-US New Partnership Fund (PMI2).

C. Venkataraman (✉)
Mathematics Institute, Zeeman Building, University of Warwick, Coventry, CV4 7AL, UK
e-mail: c.venkataraman@warwick.ac.uk

O. Lakkis · A. Madzvamuse
Department of Mathematics, University of Sussex, Pev III, Brighton, BN1 9QH, UK
e-mail: O.Lakkis@sussex.ac.uk; A.Madzvamuse@sussex.ac.uk

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_8, © Springer-Verlag Berlin Heidelberg 2013

where Ω_t is a simply connected bounded continuously deforming domain with respect to t, with Lipschitz boundary $\partial\Omega_t$ at time $t \in [0, T]$. The vector of nonlinear coupling terms $\boldsymbol{f} := (f_1, \ldots, f_m)^{\mathsf{T}}$ is assumed to be locally Lipschitz-continuous, $\boldsymbol{D} := (D_1, \ldots, D_m)^{\mathsf{T}}$ is a vector of strictly positive diffusion coefficients, $\boldsymbol{a} = (a_1, \ldots, a_d)^{\mathsf{T}}$ is a flow velocity generated by the evolution of the domain, and the initial data $\boldsymbol{u}^0(\boldsymbol{x})$ is a bounded vector valued function. Systems of this form arise in the theory of biological pattern formation [2].

Let $\hat{\Omega}$ be a simply connected time-independent reference domain with Lipschitz boundary. We assume there exists a time-differentiable family of C^∞-diffeomorphisms $\mathscr{A} : \hat{\Omega} \times [0, T] \to \Omega_t$ such that at each instant $t \in [0, T]$ and for each $\boldsymbol{x} \in \Omega_t$ there exists a $\boldsymbol{\xi} \in \hat{\Omega}$ such that

$$\mathscr{A}(\boldsymbol{\xi}, t) = \boldsymbol{x}. \tag{2}$$

Based on the derivation presented in [3] we introduce a *weak formulation* associated with Problem (1) on the *time-independent reference domain*. The problem is to find $\hat{u}_i \in \mathrm{L}_2\big(0, T; \mathrm{H}^1(\hat{\Omega})\big)$ with $\partial_t \hat{u}_i \in \mathrm{L}_2\big(0, T; \mathrm{H}^1(\hat{\Omega})'\big)$ such that for all $t \in (0, T]$,

$$\left\langle \partial_t (J \hat{u}_i), \hat{\chi} \right\rangle_{\hat{\Omega}} + \left\langle D_i J \boldsymbol{K} \nabla \hat{u}_i, \boldsymbol{K} \nabla \hat{\chi} \right\rangle_{\hat{\Omega}} = \left\langle J f_i(\hat{\boldsymbol{u}}), \hat{\chi} \right\rangle_{\hat{\Omega}}, \quad \forall \hat{\chi} \in \mathrm{H}^1(\hat{\Omega}). \tag{3}$$

Here $\mathrm{H}^1(\hat{\Omega})'$ is the dual of $\mathrm{H}^1(\hat{\Omega})$ equipped with the norm

$$\|v\|_{\mathrm{H}^1(\hat{\Omega})'} := \sup_{w \in \mathrm{H}^1(\hat{\Omega}), w \neq 0} \frac{\langle v \mid w \rangle}{\|w\|_{\mathrm{H}^1(\hat{\Omega})}}, \tag{4}$$

where $< \cdot | \cdot >$ denotes the duality pairing between H^1 and its dual. The matrix $\boldsymbol{K}$ and J are the inverse and determinant of the Jacobian of the diffeomorphism $\mathscr{A}$ respectively.

2 A Posteriori Error Estimates

Here we state a Theorem, and the associated Assumptions under which the Theorem holds, that shows the error in the *semidiscrete* scheme can be bounded by a computable a posteriori error estimator, based on the element *residual*. Our strategy to derive an a posteriori error estimate is similar to that employed by Kruger et al. [4]. We use energy arguments to show the residual is an upper bound for the error and the analysis is similar to the a priori case we have considered elsewhere [3]. For the details of the proofs we refer to [2].

We start by stating the semidiscrete scheme, find $\hat{u}_i^h : [0, T] \to \hat{\mathbb{V}}$, such that for $i = 1, \ldots, m$,

$$\begin{cases} \left\langle \partial_t (J \hat{u}_i^h), \hat{\Phi} \right\rangle_{\hat{\Omega}} + \left\langle D_i J \boldsymbol{K} \nabla \hat{u}_i^h, \boldsymbol{K} \nabla \hat{\Phi} \right\rangle_{\hat{\Omega}} = \left\langle J \widetilde{f}_i(\hat{\boldsymbol{u}}^h), \hat{\Phi} \right\rangle_{\hat{\Omega}} & \forall \hat{\Phi} \in \hat{\mathbb{V}} \text{ and } t \in (0, T], \\ \hat{u}_i^h(0) = \Lambda^h \hat{u}_i^0, \end{cases} \tag{5}$$

where $\hat{\mathbb{V}}$ is a standard FE space made up of piecewise polynomial functions and $\Lambda^h : H^1(\hat{\Omega}) \to \hat{\mathbb{V}}$ is the Lagrange interpolant.

Assumption 1 (Applicability of the mean value theorem). We assume that

$$\left\| \boldsymbol{f}' \right\|_{\mathrm{L}_\infty\left(\mathrm{range}(\hat{\boldsymbol{u}})\right)} + \left\| \boldsymbol{f}' \right\|_{\mathrm{L}_\infty\left(\mathrm{range}(\hat{\boldsymbol{u}}^h)\right)} < \tilde{C}. \tag{6}$$

Note this assumption is satisfied if we assume a global smallness condition on the mesh-size [3] and that the continuous problem is well posed [5].

We define the error in the semidiscrete scheme

$$\hat{\boldsymbol{e}}(t) := \hat{\boldsymbol{u}}^h(t) - \hat{\boldsymbol{u}}(t), \text{ for } t \in [0, T]. \tag{7}$$

Assumption 2 (Dominant energy norm error). Since we are primarily interested in problems posed on long time intervals, we wish to circumvent the use of Gronwall's inequality. To this end we assume that the error in the $\mathrm{L}_2(0, T; \mathrm{L}_2(\hat{\Omega}))$ norm converges faster than the error in the $\mathrm{L}_2(0, T; \mathrm{H}^1(\hat{\Omega}))$ norm. We assume there exists $C^\dagger, C > 0$ and $r \in (0, 1]$ independent of the mesh-size $\hat{h}$ such that

$$\int_0^T \|\hat{\boldsymbol{e}}\|^2_{\mathrm{L}_2(\hat{\Omega})^m} \le C^\dagger \hat{h}^{2r} \sum_{i=1}^m \int_0^T \|\nabla \hat{e}_i\|^2_{\mathrm{L}_2(\hat{\Omega})} \text{ thus } \int_0^T \|\boldsymbol{e}\|^2_{\mathrm{L}_2(\Omega_t)^m} \le C \hat{h}^{2r} \sum_{i=1}^m \int_0^T \|\nabla e_i\|^2_{\mathrm{L}_2(\Omega_t)}, \tag{8}$$

where we have used the equivalence of norms between the reference and evolving domains.

We note assumptions of this type have been used previously by Kruger et al. [4] and Medina et al. [6] to obtain a posteriori estimates for quasilinear reaction-diffusion and nonlinear convection-diffusion problems.

We start by introducing the *residual* $\hat{R}_i \in \mathrm{H}^1(\hat{\Omega})'$ (the dual of $\mathrm{H}^1(\hat{\Omega})$) a.e. in $[0, T]$ which satisfies

$$\left\langle \hat{R}_i \mid \hat{\chi} \right\rangle := \left\langle \partial_t (J \hat{u}^h{}_i) - D_i \nabla \cdot \left(J \boldsymbol{K} \boldsymbol{K}^\mathsf{T} \nabla \hat{u}^h_i \right) - J f_i(\hat{\boldsymbol{u}}^h) \mid \hat{\chi} \right\rangle \quad \forall \hat{\chi} \in \mathrm{H}^1(\hat{\Omega}). \tag{9}$$

We now show the residual is an upper bound for the error.

Proposition 1 (Upper bound for the error). *Suppose Assumptions 1 and 2 hold. Let $\hat{\boldsymbol{u}}$ satisfy (3) and let the error $\hat{e}_i$ and the residual $\hat{R}_i$ be as in (7) and (9) respectively. If the mesh-size satisfies a smallness condition (see [2] for details), then*

$$\|\boldsymbol{e}(T)\|^2_{\mathrm{L}_2(\Omega_t)^m} + \sum_{i=1}^m D_i \int_0^T \|\nabla e_i\|^2_{\mathrm{L}_2(\Omega_t)} \le \|\boldsymbol{e}(0)\|^2_{\mathrm{L}_2(\Omega_0)^m} + 2 \sum_{i=1}^m \int_0^T \left\langle \hat{R}_i \mid \hat{e}_i \right\rangle. \tag{10}$$

We now introduce a concrete error estimator. For simplicity we restrict the discussion to the case of $\mathbb{P}^1$ elements and regular triangulations, the results may be straightforwardly generalised to higher order elements. For any simplex s of the triangulation $\hat{\mathscr{T}}$ we denote by $\hat{h}_s$ the diameter of s. Let E_s be the set of three edges of s. Let E_i be an edge on the interior of $\hat{\Omega}$, with outward pointing (with respect to s) normal $\boldsymbol{\nu}$. We denote by $[\![\nabla\phi\cdot\boldsymbol{\nu}]\!]_{E_i}$ the jump of $\nabla\phi\cdot\boldsymbol{\nu}$ across the edge E_i. For boundary edges we take $[\![\nabla\phi\cdot\boldsymbol{\nu}]\!]=2\nabla\phi\cdot\boldsymbol{\nu}$. The local error indicator is given by

$$(\hat{\eta}_{i|s})^2 := \hat{h}_s^2 \left\| \partial_t (J\hat{u}_i^h) - D_i \nabla\cdot\left(J\boldsymbol{K}\boldsymbol{K}^{\mathsf{T}}\nabla\hat{u}_i^h\right) - J f_i(\hat{\boldsymbol{u}}^h)\right\|^2_{\mathrm{L}_2(s)} + \frac{1}{2}\sum_{e\in E_s}|e|\left\| D_i \left[\!\left[J\boldsymbol{K}\boldsymbol{K}^{\mathsf{T}}\nabla\hat{u}_i^h\cdot\boldsymbol{\nu}\right]\!\right]\right\|^2_{\mathrm{L}_2(e)}. \tag{11}$$

Proposition 2 (Residual bound). *Let R_i and $\hat{\eta}_i$, $i=1,\dots,m$ be defined by (9) and (11) respectively. There exists a $C>0$ that depends only on the shape regularity of the triangulation $\hat{\mathscr{T}}$ such that for $i=1,\dots,m$,*

$$\left|\int_0^T \left\langle \hat{R}_i \mid \hat{\chi}\right\rangle\right| \le C\int_0^T \left(\sum_{s\in\hat{\mathscr{T}}}(\hat{\eta}_{i|s})^2\right)^{1/2}\int_0^T \left\|\hat{\chi}\right\|_{\mathrm{H}^1(\hat{\Omega})} \quad \forall\hat{\chi}\in \mathrm{L}_2([0,T];\mathrm{H}^1(\hat{\Omega})). \tag{12}$$

To complete the bound of the error by the estimator, we make an assumption about the error in the approximation of the initial data.

Assumption 3 (Dominated initial error). We assume that the initial error in the $\mathrm{L}_2(\hat{\Omega})$ norm converges faster than the error in the $\mathrm{L}_2(0,T;\mathrm{H}^1(\hat{\Omega}))$. We assume there exists $C>0$ and $r\in(0,1]$ both independent of the mesh-size $\hat{h}$ such that

$$\left\|\hat{\boldsymbol{e}}(0)\right\|^2_{\mathrm{L}_2(\hat{\Omega})} \le C\hat{h}^{2r}\sum_{i=1}^m\int_0^T \left\|\nabla\hat{e}_i\right\|^2_{\mathrm{L}_2(\hat{\Omega})}. \tag{13}$$

Theorem 1 (A posteriori error estimate for the semidiscrete scheme). *Let Assumptions 1–3 hold. Let the error $\hat{e}_i$ and the estimator $\hat{\eta}_i$, $i=1,\dots,m$ be defined by (7) and (11) respectively. If the mesh-size is sufficiently small, for some $C>0$, we have*

$$\sum_{i=1}^m D_i\int_0^T \|\nabla e_i\|^2_{\mathrm{L}_2(\Omega_t)} \le \sum_{i=1}^m C\int_0^T\sum_{s\in\hat{\mathscr{T}}}(\hat{\eta}_{i|s})^2. \tag{14}$$

Since the estimator $\hat{\boldsymbol{\eta}}$ is an upper bound for the error, we use it to drive a space-adaptive scheme. To ensure the efficiency of the adaptive scheme, we would have to show the estimator was also a lower bound for the error and we leave this extension for future work.

3 Numerical Results

Here we reconcile our theoretical results with numerical computations. We start by presenting a time discretisation of the *semidiscrete* scheme (5), we discretise in time using a modified implicit Euler method [7], in which the reaction terms are treated semi-implicitly while the diffusive terms are treated fully implicitly: find $(\hat{U}_i^h)^n \in \hat{\mathbb{V}}^n$, such that for $i = 1, \ldots, m$, $n = 1, \ldots, N$,

$$\begin{cases} \left\langle \frac{1}{\tau}\bar{\partial}\left[J\hat{(}U_i^h)\right]^n, \hat{\Phi}^n \right\rangle_{\hat{\Omega}} + \left\langle D_i[J\boldsymbol{K}\nabla\hat{(}U_i^h)]^n, [\boldsymbol{K}\nabla\hat{\Phi}]^n \right\rangle_{\hat{\Omega}} = \left\langle J^n\widetilde{f}_i((\hat{U}_i^h)^n, (\hat{U}_i^h)^{n-1}), \hat{\Phi}^n \right\rangle_{\hat{\Omega}} \forall \hat{\Phi} \in \hat{\mathbb{V}}^n, \\ (\hat{U}_i^h)^0 = \Lambda^h \hat{u}_i^0, . \end{cases} \tag{15}$$

The adaptive algorithm we consider is based on the equidistribution marking strategy [8, Algorithm 1.19, p. 45], where elements are marked for refinement and coarsening with the goal of equidistributing the estimator value over all mesh elements. The marking strategy takes two parameters: the tolerance of the adaptive algorithm tol and a parameter $\theta \in (0, 1)$. At each timestep elements are marked for refinement according to the following algorithm:

3.1 Equidistribution Strategy (Refinement)

Start with $\hat{\mathscr{T}}_0^n$ the initial triangulation at time n, tolerance tol and parameter θ
$k := 0$
solve the discrete linear problem on the mesh $\hat{\mathscr{T}}_k^n$
compute global error estimator $\hat{\boldsymbol{\eta}}$ and local error indicators $\hat{\boldsymbol{\eta}}_{|s}$
while $\hat{\boldsymbol{\eta}} > tol$ **do**
  for all $s \in \hat{\mathscr{T}}_k^n$ **do**
    if $\hat{\boldsymbol{\eta}}_{|s} > \theta * tol/N$ {where N is the number of elements of the triangulation}
    then
      mark s for refinement {elements are also marked for coarsening at this stage}
    end if
  end for
  adapt mesh $\hat{\mathscr{T}}_k^n$ to give $\hat{\mathscr{T}}_{k+1}^n$
  $k := k + 1$
  solve the discrete linear problem on the mesh $\hat{\mathscr{T}}_k^n$
  Compute global error estimator $\hat{\boldsymbol{\eta}}$ and local error indicators $\hat{\boldsymbol{\eta}}_{|s}$
end while

Elements are marked for coarsening in a similar way to the above, the difference being that if the local error indicator plus a coarsening indicator is less than a given tolerance on an element then the element is marked for coarsening [8, p. 48].

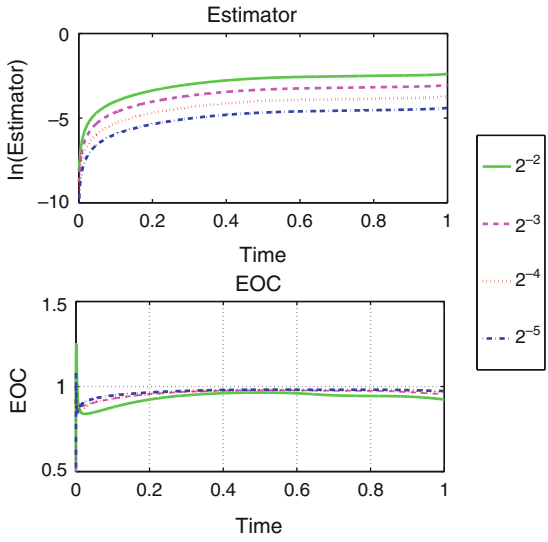


Fig. 1 The log of the $L_2([0, T])$ norm of the estimator $\hat{\eta}$ (cf. (11)) and the EOC of the estimator against time for a benchmark computation. The legend indicates the mesh-size $\hat{h}$ for each simulation

For numerical testing, we consider Problem (1) equipped with the Schnakenberg kinetics [9]:

$$f_1(\boldsymbol{u}) = \gamma\left(k_1 - u_1 + u_1^2 u_2\right) \text{ and } f_2(\boldsymbol{u}) = \gamma\left(k_2 - u_1^2 u_2\right), \tag{16}$$

where $0 < \gamma, k_1, k_2 < \infty$. The details of the implementation of the scheme are described elsewhere [3]. We consider a simulation of an RDS equipped with the Schnakenberg kinetics, with parameter values $\boldsymbol{D} = (1, 10)^{\mathsf{T}}$, $k_1 = 0.1$, $k_2 = 0.9$, $\gamma = 1$, adding source terms such that the exact solution is known, on a domain with evolution of the form

$$\mathscr{A}(\boldsymbol{\xi}, t) = \boldsymbol{\xi}(1 + \sin(\pi t)), \quad \boldsymbol{\xi} \in [0, 1]^2, t \in [0, 1]. \tag{17}$$

We used a sufficiently small timestep such that the error due to the time discretisation is negligible. The estimator values and EOC for a series of refinements are plotted in Fig. 1 and we observe an EOC of 1, as expected for $\mathbb{P}^1$ elements providing numerical evidence for Theorem 1.

We next present results for the Schnakenberg kinetics, with parameter values $\boldsymbol{D} = (0.01, 1)^{\mathsf{T}}$, $k_1 = 0.1$, $k_2 = 0.9$, $\gamma = .1$, where no exact solution is known on a domain with evolution of the form

$$\mathscr{A}(\boldsymbol{\xi}, t) = \boldsymbol{\xi}(1 + 9\sin(\pi t/1000)), \quad \boldsymbol{\xi} \in [0, 1]^2 \text{ and } t \in [0, 1000]. \tag{18}$$

We consider an adaptive scheme based on the equidistribution marking strategy with parameters $\theta = 0.8$, $tol = 10^{-4}$ and a fixed timestep of 10^{-2}. Figures 2 and 3 show

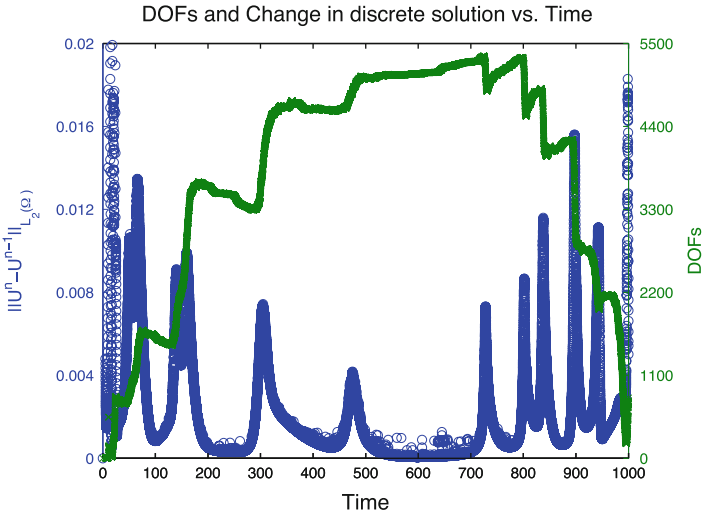


Fig. 2 The number of DOFs (*green crosses*) and the change in discrete solution (*blue circles*) vs. time for the Schnakenberg kinetics on a domain with evolution of the form (18). The number of DOFs appears positively correlated with the domain size. Bifurcations in the discrete solution correspond to spikes in the change in discrete solution. Spot-splitting bifurcations lead to increases in DOFs, while spot-annihilation or -merging results in decreases in DOFs

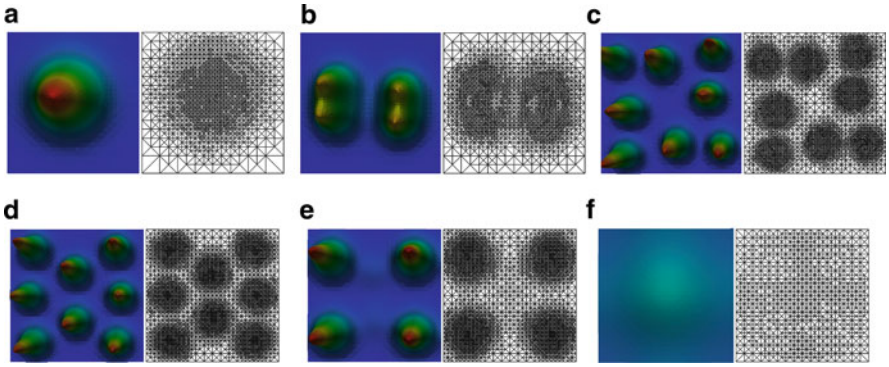


Fig. 3 Snapshots of the discrete activator (u_1) profile for the Schnakenberg kinetics on the reference domain, under adaptive mesh refinement and domain evolution of the form (18). (**a**) $t = 50$. (**b**) $t = 160$. (**c**) $t = 380$. (**d**) $t = 700$. (**e**) $t = 820$. and (**f**) $t = 1{,}000$

the evolution of the degrees of freedom (DOFs) and the change in discrete solution and snapshots of the activator profiles (on the reference domain) respectively. The number of DOFs appears positively correlated with the domain size. The mesh is also well refined around the patterns during the evolution, illustrating the benefits of adaptive mesh refinement.

We finish with an application to the case where the evolving domain is an *evolving surface* embedded in $\mathbb{R}^3$ that is diffeomorphic to a time-independent planar domain. We have derived the model equations and corresponding finite element method elsewhere [10] and thus only briefly state the details. The model for an RDS posed on an evolving surface is of the form: find u_i, functions from Γ_t into $\mathbb{R}$, such that for $i = 1, \ldots, m$, u_i satisfies

$$\begin{cases} \partial_t u_i(\boldsymbol{x},t) + [\boldsymbol{a}\cdot\nabla u_i](\boldsymbol{x},t) - D_i\Delta_{\Gamma_t} u_i(\boldsymbol{x},t) + \left[u_i \nabla_{\Gamma_t}\cdot\boldsymbol{a}\right](\boldsymbol{x},t) = f_i\left(\boldsymbol{u}(\boldsymbol{x},t)\right), & \boldsymbol{x}\in\Gamma_t, t\in(0,T], \\ [\boldsymbol{v}\cdot\nabla_{\Gamma_t} u_i](\boldsymbol{x},t) = 0, & \boldsymbol{x}\in\partial\Gamma_t, t>0, \\ u_i(\boldsymbol{x},0) = u_i^0(\boldsymbol{x}), & \boldsymbol{x}\in\Gamma_0, \end{cases} \tag{19}$$

here the Cartesian gradient and Laplacian that appear in (1) are replaced by the surface (tangential) gradient and Laplace-Beltrami operator. We assume the surface Γ_t admits an *orthogonal parameterisation* to a planar domain which we denote by

$$\mathscr{A} : \hat{\Omega} \subset \mathbb{R}^2 \times [0,T] \to \Gamma_t \subset \mathbb{R}^3. \tag{20}$$

Under similar Assumptions to those made in the planar case (see [10] for details) the corresponding weak formulation on the reference domain $\hat{\Omega}$ is given by (3) where the matrix $\boldsymbol{K}$ and the determinant of the Jacobian J are given by

$$\boldsymbol{K} = \begin{bmatrix} 1/|\partial_1\mathscr{A}| & 0 \\ 0 & 1/|\partial_2\mathscr{A}| \end{bmatrix} \quad \text{and} \quad J = |\partial_1\mathscr{A}|\,|\partial_2\mathscr{A}|\,. \tag{21}$$

We consider an example with the Schnakenberg kinetics (16), with parameter values $\boldsymbol{D} = (0.01, 1)^{\mathsf{T}}$, $k_1 = 0.1$, $k_2 = 0.9$, $\gamma = 1$, where no exact solution is known on a domain with evolution of the form

$$\mathscr{A}_1(\boldsymbol{\xi},t) = \xi_1,\ \mathscr{A}_2(\boldsymbol{\xi},t) = \xi_2,\ \mathscr{A}_3(\boldsymbol{\xi},t) = 4\sin(\pi t/500)(\xi_1-\xi_2)^4 \quad \boldsymbol{\xi}\in[0,1]^2, t\in[0,500]. \tag{22}$$

We once again consider an adaptive scheme based on the equidistribution marking strategy with parameters $\theta = 0.8, tol = 10^{-3}$ and a fixed timestep of 10^{-2}. Figure 4 shows snapshots of the activator profiles (on the surface and on the reference domain) and the mesh of the reference domain. As the surface evolves, we observe the emergence of a large number of spots with small radii in the top left and bottom right hand corners of the domain (where curvature is large and growth is fastest) with annihilation of these spots as the domain contracts. The results clearly illustrate the influence of growth and curvature on pattern formation. The adaptive scheme appears to resolve the solution profiles and the mesh is well refined around the spots on the reference domain, capturing both the small radii spots that develop in the Northwest and Southeast corners and the large radii spots that develop elsewhere.

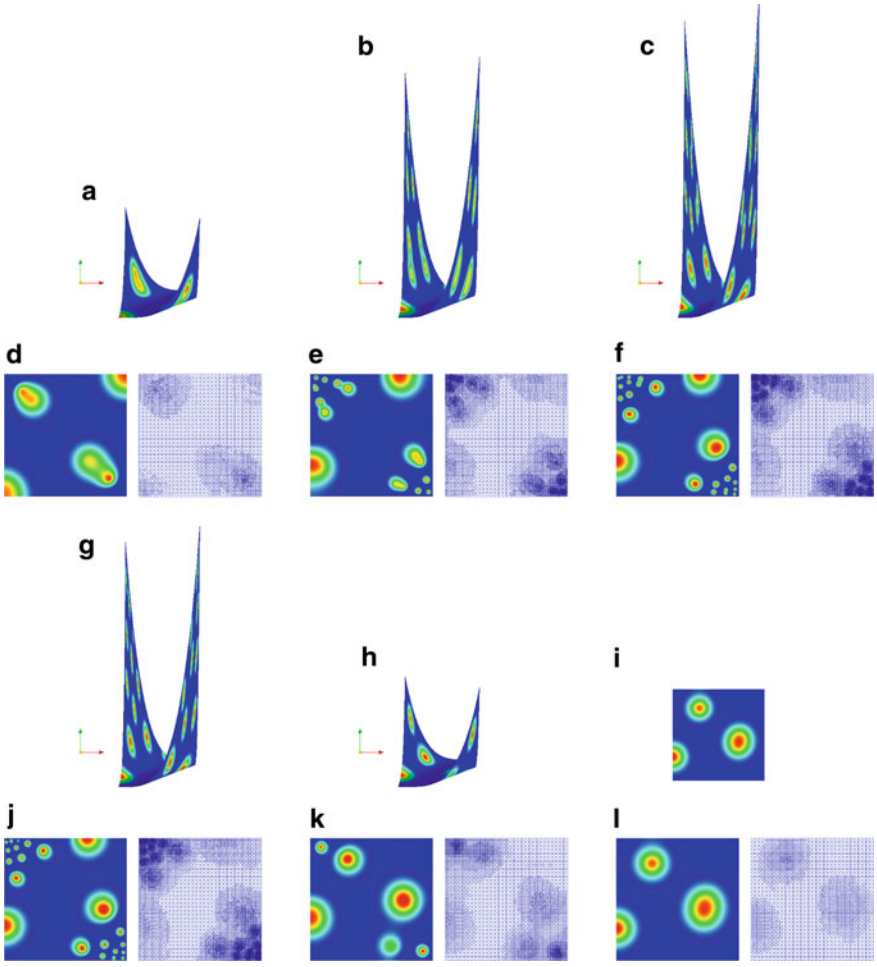


Fig. 4 Snapshots of the discrete activator (u_1) profile for the Schnakenberg kinetics on the evolving surface and on the reference domain together with the mesh, under adaptive mesh refinement and domain evolution of the form (22). (**a**) $t = 50$ (surface). (**b**) $t = 150$ (surface). (**c**) $t = 250$ (surface). (**d**) $t = 50$ (reference and mesh). (**e**) $t = 150$ (reference and mesh). (**f**) $t = 250$ (reference and mesh). (**g**) $t = 350$ (surface). (**h**) $t = 450$ (surface). (**i**) $t = 500$ (surface). (**j**) $t = 350$ (reference and mesh). (**k**) $t = 450$ (reference and mesh). (**l**) $t = 500$ (reference and mesh)

Finally, we remark that we have also considered space-time adaptive schemes based on an heuristic error indicator for the time adaptivity which appear to give dramatic improvements in efficiency [2].

References

1. A. Madzvamuse. *A Numerical Approach to the Study of Spatial Pattern Formation*. DPhil thesis, University of Oxford, 2000.
2. C. Venkataraman. *Reaction-diffusion systems on evolving domains with applications to the theory of biological pattern formation*. DPhil thesis, University of Sussex, June 2011. URL http://sro.sussex.ac.uk/6908/.
3. O. Lakkis, A. Madzvamuse, and C. Venkataraman. Implicit-explicit timestepping with finite element approximation of reaction-diffusion systems on evolving domains. Technical report, March 2012. URL http://adsabs.harvard.edu/abs/2011arXiv1111.5052L.
4. O. Kruger, M. Picasso, and JF Scheid. A posteriori error estimates and adaptive finite elements for a nonlinear parabolic problem related to solidification. *Computer Methods in Applied Mechanics and Engineering*, 192(5–6):535–558, 2003.
5. Chandrasekhar Venkataraman, Omar Lakkis, and Anotida Madzvamuse. Global existence for semilinear reaction–diffusion systems on evolving domains. *Journal of Mathematical Biology*, 64:41–67, 2012. ISSN 0303-6812. URL http://dx.doi.org/10.1007/s00285-011-0404-x.
6. J. Medina, M. Picasso, and J. Rappaz. Error estimates and adaptive finite elements for nonlinear diffusion-convection problems. *Mathematical Models and Methods in Applied Sciences*, 6(5): 689–712, 1996. ISSN 0218-2025.
7. A. Madzvamuse. A modified backward euler scheme for advection-reaction-diffusion systems. *Mathematical Modeling of Biological Systems, Volume I*, pages 183–189, 2007.
8. A. Schmidt and K.G. Siebert. *Design of adaptive finite element software: The finite element toolbox ALBERTA*. Springer Verlag, 2005.
9. R. Lefever and I. Prigogine. Symmetry-breaking instabilities in dissipative systems II. *J. chem. Phys*, 48:1695–1700, 1968.
10. C. Venkataraman, T. Sekimura, E.A. Gaffney, P.K. Maini, and A. Madzvamuse. Modeling parr-mark pattern formation during the early development of amago trout. *Phys. Rev. E*, 84: 041923, Oct 2011. doi: 10.1103/PhysRevE.84.041923.

Moment-Based Boundary Conditions for Lattice Boltzmann Magnetohydrodynamics

P.J. Dellar

Abstract We present a moment-based approach for implementing boundary conditions in a lattice Boltzmann formulation of magnetohydrodynamics. Hydrodynamic quantities are represented using a discrete set of distribution functions that evolve according to a cut-down form of Boltzmann's equation from continuum kinetic theory. Electromagnetic quantities are represented using a set of vector-valued distribution functions. The nonlinear partial differential equations of magnetohydrodynamics are thus replaced by two constant-coefficient hyperbolic systems in which all nonlinearities are confined to algebraic source terms. Further discretising these systems in space and time leads to efficient and readily parallelisable algorithms. However, the widely used bounce-back boundary conditions place no-slip boundaries approximately half-way between grid points, with the precise position being a function of the viscosity and resistivity. Like most lattice Boltzmann boundary conditions, bounce-back is inspired by a discrete analogue of the diffuse and specular reflecting boundary conditions from continuum kinetic theory. Our alternative approach using moments imposes no-slip boundary conditions precisely at grid points, as demonstrated using simulations of Hartmann flow between two parallel planes.

1 Introduction

The lattice Boltzmann approach to computational fluid dynamics is based on a disrete analogue of Boltzmann's equation from the kinetic theory of gases [4, 5, 16, 18]. The particle velocity $\boldsymbol{\xi}$ is restricted to a discrete set $\boldsymbol{\xi}_0, \ldots, \boldsymbol{\xi}_N$. The corresponding distribution functions $f_i(\mathbf{x}, t)$ evolve according to the discrete Boltzmann equation

P.J. Dellar (✉)
Mathematical Institute, University of Oxford, 24–29 St Giles', Oxford, OX1 3LB, UK
e-mail: dellar@maths.ox.ac.uk

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_9,

$$\partial_t f_i + \boldsymbol{\xi}_i \cdot \nabla f_i = -\sum_{j=0}^{N} \Omega_{ij} \left(f_j - f_j^{(0)} \right). \tag{1}$$

Hydrodynamic quantities such as the fluid density ρ, velocity $\mathbf{u}$, and momentum flux $\boldsymbol{\Pi}$ are given by moments of the f_i,

$$\rho = \sum_{i=0}^{N} f_i, \quad \rho\mathbf{u} = \sum_{i=0}^{N} \boldsymbol{\xi}_i f_i, \quad \boldsymbol{\Pi} = \sum_{i=0}^{N} \boldsymbol{\xi}_i \boldsymbol{\xi}_i f_i, \quad \mathbf{Q} = \sum_{i=0}^{N} \boldsymbol{\xi}_i \boldsymbol{\xi}_i \boldsymbol{\xi}_i f_i. \tag{2}$$

These sums replace the integrals over $\boldsymbol{\xi}$ in classical kinetic theory [4]. The equilibrium distributions $f_j^{(0)}(\rho, \mathbf{u})$ and collision matrix Ω_{ij} are chosen so that slowly varying solutions of the moment hierarchy

$$\partial_t \rho + \nabla\cdot(\rho\mathbf{u}) = 0, \quad \partial_t(\rho\mathbf{u}) + \nabla\cdot\boldsymbol{\Pi} = 0, \quad \partial_t \boldsymbol{\Pi} + \nabla\cdot\mathbf{Q} = -\frac{1}{\tau}\left(\boldsymbol{\Pi} - \boldsymbol{\Pi}^{(0)} \right) \tag{3}$$

obtained from (1) satisfy the isothermal Navier–Stokes equations on timescales much longer than the timescale τ associated with collisions. We obtain the Euler equations with constant sound speed c_s by setting $\boldsymbol{\Pi}^{(0)} = c_\mathrm{s}^2 \rho \mathbf{I} + \rho\mathbf{u}\mathbf{u}$, where $\mathbf{I}$ is the identity tensor. The first correction to the momentum flux is given by

$$\boldsymbol{\Pi} = \boldsymbol{\Pi}^{(0)} - \tau c_\mathrm{s}^2 \rho \big((\nabla\mathbf{u}) + (\nabla\mathbf{u})^\mathsf{T} \big), \tag{4}$$

corresponding to a Newtonian viscous stress with dynamic viscosity $\mu = \tau c_\mathrm{s}^2 \rho$.

The constant coefficient hyperbolic system (1) is readily discretised by integration along characteristics [12], or by splitting into separate advection and collision steps [8], to obtain the fully discrete system [7]

$$\overline{f}_i(\mathbf{x} + \boldsymbol{\xi}_i \Delta t, t + \Delta t) = \overline{f}_i(\mathbf{x}, t) - \Delta t \sum_{j=0}^{N} \overline{\Omega}_{ij} \left(\overline{f}_j(\mathbf{x}, t) - f_j^{(0)}(\mathbf{x}, t) \right), \tag{5}$$

under the change of variables $\overline{f}_i = f_i + \frac{1}{2}\Delta t \sum_{j=0}^{N} \Omega_{ij} \left(f_j - f_j^{(0)} \right)$. The discrete collision matrix is $\overline{\boldsymbol{\Omega}} = \left(\mathbf{I} + \frac{1}{2}\Delta t \boldsymbol{\Omega} \right)^{-1} \boldsymbol{\Omega}$. These formulae reduce to the standard redefinition of the collision time from τ to $\tau + \Delta t/2$ for the single-relaxation-time collision operator $\Omega_{ij} = \tau^{-1}\delta_{ij}$.

2 The Magnetic Field

Magnetohydrodynamics describes the interaction between electrically conducting fluids and magnetic fields [3, 14]. The Lorentz force exerted by a magnetic field $\mathbf{B}$ may be expressed as the divergence of the Maxwell stress $\frac{1}{2}|\mathbf{B}|^2 \mathbf{I} - \mathbf{B}\mathbf{B}$.

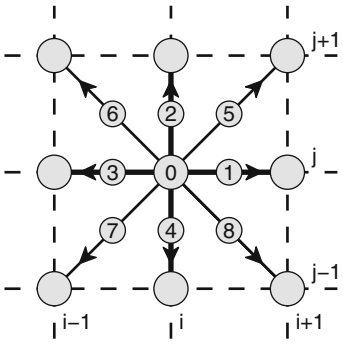


Fig. 1 Sketch of the discrete velocities $\boldsymbol{\xi}_0, \ldots, \boldsymbol{\xi}_8$ arranged on an integer lattice. Only the velocities $\boldsymbol{\xi}_0, \ldots, \boldsymbol{\xi}_4$ shown with *thicker arrows* are used for the magnetic distribution functions

These two terms give an isotropic magnetic pressure and a tension directed along field lines. The Lorentz force is thus readily incorporated into lattice Boltzmann hydrodynamics by choosing the equilibrium momentum flux to be [6]

$$\boldsymbol{\Pi}^{(0)} = c_s^2 \rho \mathbf{I} + \rho \mathbf{u}\mathbf{u} + \tfrac{1}{2}|\mathbf{B}|^2 \mathbf{I} - \mathbf{B}\mathbf{B}. \tag{6}$$

Suitable two-dimensional equilibria (in units where $c_s = 1/3$) are given by

$$f_i^{(0)} = w_i \left[\rho \left(2 - \tfrac{3}{2}|\boldsymbol{\xi}_i|^2 \right) + 3\rho \mathbf{u} \cdot \boldsymbol{\xi}_i + \tfrac{9}{2} \boldsymbol{\Pi}^{(0)} : \boldsymbol{\xi}_i \boldsymbol{\xi}_i - \tfrac{3}{2} \operatorname{Tr} \boldsymbol{\Pi}^{(0)} \right]. \tag{7}$$

The discrete velocities $\boldsymbol{\xi}_0, \ldots, \boldsymbol{\xi}_8$ form an integer square lattice in these units, see Fig. 1. The corresponding weights are $w_0 = 4/9$, $w_{1,2,3,4} = 1/9$, and $w_{5,6,7,8} = 1/36$. The expressions (7) reduce to the standard D2Q9 equilibria [16] when $\mathbf{B} = 0$.

The magnetic field evolves through Faraday's law $\partial_t \mathbf{B} + \nabla \times \mathbf{E} = 0$, where the electric field $\mathbf{E}$ is given by Ohm's law $\mathbf{E} + \mathbf{u} \times \mathbf{B} = \eta \nabla \times \mathbf{B}$ in resistive magnetohydrodynamics. Faraday's law cannot be derived from a kinetic equation of the form (1), because the vector $\rho \mathbf{u}$ evolves through the divergence of the symmetric tensor $\boldsymbol{\Pi}$ in (3). By contrast, $\partial_t \mathbf{B} = -\nabla \cdot \boldsymbol{\Lambda}$ evolves through the divergence of an antisymmetric tensor whose components $\Lambda_{\alpha\beta} = -\epsilon_{\alpha\beta\gamma} E_\gamma$ are formed by contracting the electric field with the alternating tensor.

Instead, we represent the magnetic field as $\mathbf{B} = \sum_i \mathbf{g}_i$ using a set of vector-valued distribution functions $\mathbf{g}_i$ that evolve according to the vector Boltzmann equation [6]

$$\partial_t \mathbf{g}_i + \boldsymbol{\xi}_i \cdot \nabla \mathbf{g}_i = -\frac{1}{\tau_m} \left(\mathbf{g}_i - \mathbf{g}_i^{(0)} \right), \tag{8}$$

with the equilibrium distributions

$$\mathbf{g}_i^{(0)} = W_i \big(\mathbf{B} - 3\, \boldsymbol{\xi}_i \times (\mathbf{u} \times \mathbf{B}) \big). \tag{9}$$

The magnetic weights are $W_0 = 1/3$ and $W_{1,2,3,4} = 1/6$. The four diagonal velocities $\boldsymbol{\xi}_5, \ldots, \boldsymbol{\xi}_8$ are not needed for the magnetic distribution functions.

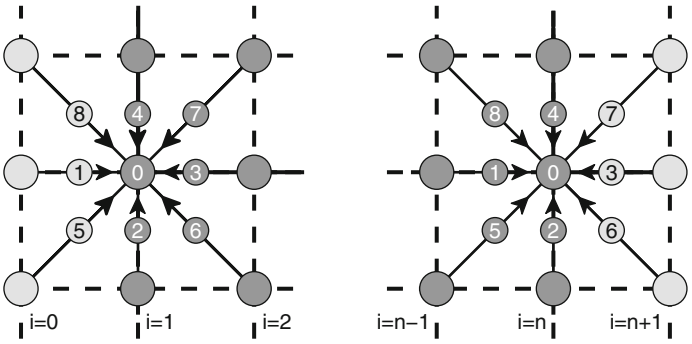


Fig. 2 The boundary conditions must supply values for the incoming distributions f_1, f_5, f_8 on the left boundary, and for f_3, f_6, f_7 on the right boundary

Slowly varying solutions of (8) obey the correct evolution equation for a magnetic field under resistive magnetohydrodynamics [6]

$$\partial_t \mathbf{B} = \nabla\times(\mathbf{u}\times\mathbf{B}) + \nabla\cdot(\eta\nabla\mathbf{B}), \tag{10}$$

with resistivity given by $\eta = \tau_m/3$ in the so-called lattice units with $|\boldsymbol{\xi}_{1,2,3,4}| = 1$. Discretising (8) leads to a numerical scheme analogous to (5) that is coupled to the hydrodynamic lattice Boltzmann equation through the macroscopic velocity and magnetic field at grid points. The resulting numerical scheme preserves $\nabla\cdot\mathbf{B} = 0$ to round-off error. It has been used in large-scale (up to $1{,}800^3$ grid points) simulations of three-dimensional MHD turbulence [17, 19], and to simulate liquid metal flows in cooling systems for nuclear reactors [15].

3 Boundary Conditions for Hartmann Flow

We present simulations of the MHD analogue of Poiseuille flow, known as Hartmann flow, in which a uniform pressure gradient drives a unidirectional flow along a channel spanned by an imposed uniform magnetic field. The flow stretches the imposed field to create an additional magnetic field component along the channel, and hence a Lorentz force that resists the flow. We choose axes with y directed along the channel, and x directed across the channel with the walls at $x = \pm L$. Imposing no-flux and no-slip boundary conditions corresponds to setting $\mathbf{u} = 0$ on the walls. Maxwell's equations imply continuity of the normal and tangential components of $\mathbf{B}$ at the walls, so $\mathbf{B} = (B_0, 0, 0)$ takes its external applied value [14].

Following the approach of Bennett [1, 2] we formulate boundary conditions in terms of moments of the f_i. At the left-hand boundary we must supply values for the three distributions f_1, f_5, f_8 that propagate inwards from outside the domain, as sketched in Fig. 2. This may be done by specifying values for the three moments ρu_x, ρu_y, and Π_{yy},

$$0 = \rho u_x = f_1 + f_5 + f_8 \quad - f_3 - f_6 - f_7, \tag{11a}$$

$$0 = \rho u_y = \quad f_5 - f_8 \quad + f_2 - f_4 + f_6 - f_7, \tag{11b}$$

$$\Pi^{(0)}_{yy} = \Pi_{yy} = \quad f_5 + f_8 \quad + f_2 + f_4 + f_6 + f_7. \tag{11c}$$

These three moments are chosen because they contain three linearly independent combinations of the unknowns f_1, f_5, f_8. The first two conditions (11a) and (11b) impose no-flux and no-slip boundary conditions, and the third boundary condition (11c) on the tangential stress has a more natural physical interpretation than the alternatives involving the higher moments [1, 2]. Solving this system of three linear equations determines the incoming distributions,

$$f_1 = f_2 + f_3 + f_4 + 2f_6 + 2f_7 - \Pi^{(0)}_{yy}, \tag{12a}$$

$$f_5 = -f_2 - f_6 + \tfrac{1}{2}\,\Pi^{(0)}_{yy}, \tag{12b}$$

$$f_8 = -f_4 - f_7 + \tfrac{1}{2}\,\Pi^{(0)}_{yy}. \tag{12c}$$

For this simple flow it is sufficient to take $\Pi^{(0)}_{yy} = c_s^2\rho = c_s^2$ on the boundary, since the tangential velocity and magnetic field both vanish. The fluid density ρ is uniform, and may be set equal to unity in the initial conditions. More generally, one would solve for ρ as part of the linear system by setting $\Pi^{(0)}_{yy} = c_s^2(f_0 + \cdots + f_8)$ in (11c)

A similar approach determines the incoming magnetic distributions g_{1x} and g_{1y} from the boundary conditions $B_x = B_0$ and $B_y = 0$,

$$B_x = B_0 \implies g_{x1} = B_0 - \left(g_{x0} + g_{x2} + g_{x3} + g_{x4}\right), \tag{13a}$$

$$B_y = 0 \implies g_{y1} = \quad - \left(g_{y0} + g_{y2} + g_{y3} + g_{y4}\right). \tag{13b}$$

The same approach enables f_3, f_6, f_7 and g_{x3}, g_{y3} to be determined at the right-hand wall, and we impose periodic boundary conditions in the y direction.

4 Numerical Experiments

Figure 3 shows the results of a lattice Boltzmann computation using these boundary conditions. The flow was driven by including an additional linear stress $xF\hat{\mathbf{x}}\hat{\mathbf{y}}$ into $\boldsymbol{\Pi}^{(0)}$, equivalent to a uniform body force $F\hat{\mathbf{y}}$, as in previous computations [6, 9]. The channel was taken to be the domain $|x| \le L = 0.5$ in suitable dimensionless units, with $B_0 = 1$ and $F = 1$. The resistivity was $\eta = 0.1$ and the kinematic viscosity was $\nu = \mu/\rho = 0.025$. The lattice Boltzmann simulation shown was performed with 64 points and Mach number $\mathrm{Ma} = \sqrt{3}/50$. The Mach number controls the ratio between the macroscopic fluid speed and the particle speeds, since $c_s = |\boldsymbol{\xi}_1|/\sqrt{3}$.

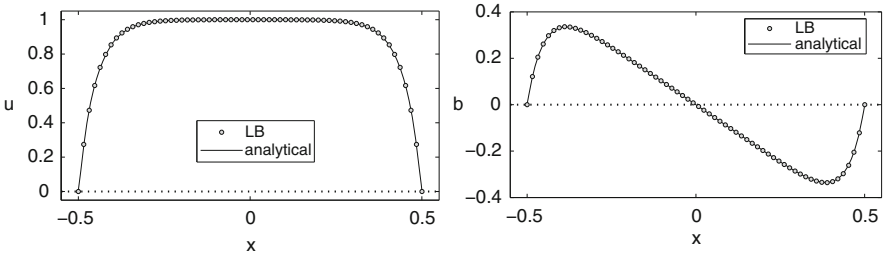


Fig. 3 Streamwise velocity and magnetic field, lattice Boltzmann (LB) computations versus the analytical solution for Hartmann number $H = 10$, $B_0 = 1$, and 64 points. The velocity and magnetic field both vanish up to round-off error at the endpoints

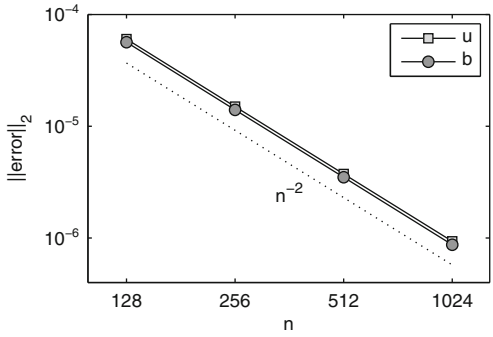


Fig. 4 Second-order convergence of the computed velocity and magnetic field towards the analytical solution (14) with increasing number of grid points n

Incompressible Hartmann flow in fluid of unit density has the exact solution [14]

$$b(x) = \frac{FL}{B_0}\left[\frac{\sinh(Hx/L)}{\sinh(H)} - \frac{x}{L}\right], \quad u(x) = \frac{FL}{B_0}\sqrt{\frac{\eta}{\nu}}\coth(H)\left[1 - \frac{\cosh(Hx/L)}{\cosh(H)}\right] \tag{14}$$

for the streamwise (y-component) velocity and magnetic field, while the spanwise magnetic field remains uniform. The Hartmann number $H = B_0 L/(\eta\nu)^{1/2}$ measures the ratio of Lorentz to viscous forces, with $H = 10$ for the parameters given. When $H \gg 1$ the streamwise velocity is nearly uniform, and the magnetic field nearly linear, outside $O(L/H)$-wide boundary layers at the walls, as shown in Fig. 3.

The streamwise velocity and magnetic field both vanish precisely at the endpoints using moment-based boundary conditions, unlike previous computations using bounce-back boundary conditions [6,9]. Figure 4 shows the discrete ℓ_2 norms of the differences between the lattice Boltzmann (LB) and analytical solutions,

$$\Delta u = \left(\frac{1}{n}\sum_{i=1}^{n} |u_{LB}(x_i) - u(x_i)|^2\right)^{1/2}, \quad \Delta b = \left(\frac{1}{n}\sum_{i=1}^{n} |b_{LB}(x_i) - b(x_i)|^2\right)^{1/2}, \tag{15}$$

for different numbers of grid points n. To achieve the expected second-order convergence rate with the simple linear forcing term in the equilibrium stress it was necessary to decrease the Mach number $\mathrm{Ma} = \sqrt{3} \times 2.56/n$ with increasing n.

5 Conclusion

Restricting the particle velocity $\boldsymbol{\xi}$ in the Boltzmann equation to a discrete set $\boldsymbol{\xi}_0, \ldots, \boldsymbol{\xi}_N$ leads to a tractable system of partial differential equations for distribution functions $f_i(\mathbf{x}, t)$. Discretising these equations in $\mathbf{x}$ and t leads to an effective tool for computational fluid dynamics. However, the widely used bounce-back boundary conditions, inspired by the diffuse and specular reflection of continuum kinetic theory, only approximate no-slip boundary conditions. The tangential velocity vanishes at a point approximately half-way between grid points, but the precise location depends on the collision rate τ unless one adopts a two-relaxation-time (TRT) collision operator with a specific ratio of relaxation times for odd and even moments [10, 11, 13]. The alternative approach of Bennett [1, 2] formulates boundary conditions for moments with direct physical interpretations, the velocity components and the tangential momentum flux. Solving the resulting linear system for the incoming distribution functions imposes no-slip and no-flux boundary conditions precisely at grid points. This approach extends easily to impose boundary conditions on a magnetic field, as shown for simulations of Hartmann flow between planar boundaries.

Acknowledgements The author's research is supported by an Advanced Research Fellowship from the Engineering and Physical Sciences Research Council [grant number EP/E054625/1].

References

1. Bennett, S.: A lattice Boltzmann model for diffusion of binary gas mixtures. Ph.D. thesis, University of Cambridge (2010). URL http://www.dspace.cam.ac.uk/handle/1810/226851.
2. Bennett, S., Asinari, P., Dellar, P.J.: A lattice Boltzmann model for diffusion of binary gas mixtures that includes diffusion slip. Int. J. Numer. Meth. Fluids. **69**, 171–189 (2012).
3. Biskamp, D.: Nonlinear Magnetohydrodynamics. Cambridge University Press, Cambridge (1993).
4. Chapman, S., Cowling, T.G.: The Mathematical Theory of Non-Uniform Gases, 3rd edn. Cambridge University Press, Cambridge (1970).
5. Chen, S., Doolen, G.D.: Lattice Boltzmann method for fluid flows. Annu. Rev. Fluid Mech. **30**, 329–364 (1998).
6. Dellar, P.J.: Lattice kinetic schemes for magnetohydrodynamics. J. Comput. Phys. **179**, 95–126 (2002).
7. Dellar, P.J.: Incompressible limits of lattice Boltzmann equations using multiple relaxation times. J. Comput. Phys. **190**, 351–370 (2003).

8. Dellar, P.J.: An interpretation and derivation of the lattice Boltzmann method using Strang splitting. Comput. Math. Applic. (2011). Published online, doi:10.1016/j.camwa.2011.08.047.
9. Dellar, P.J.: Lattice Boltzmann formulation for Braginskii magnetohydrodynamics. Computers & Fluids **46**, 201–205 (2011).
10. d'Humières, D., Ginzburg, I.: Viscosity independent numerical errors for Lattice Boltzmann models: From recurrence equations to "magic" collision numbers. Comput. Math. Applic. **58**, 823–840 (2009).
11. Ginzbourg, I., Adler, M.P.: Boundary flow condition analysis for the three-dimensional lattice Boltzmann model. J. Phys. II France **4**, 191–214 (1994).
12. He, X., Chen, S., Doolen, G.D.: A novel thermal model of the lattice Boltzmann method in incompressible limit. J. Comput. Phys. **146**, 282–300 (1998).
13. He, X.Y., Zou, Q.S., Luo, L.S., Dembo, M.: Analytic solutions of simple flows and analysis of nonslip boundary conditions for the lattice Boltzmann BGK model. J. Statist. Phys. **87**, 115–136 (1997).
14. Landau, L.D., Lifshitz, E.M.: Electrodynamics of Continuous Media. Pergamon, Oxford (1960). 2nd edition 1984.
15. Pattison, M., Premnath, K., Morley, N., Abdou, M.: Progress in lattice Boltzmann methods for magnetohydrodynamic flows relevant to fusion applications. Fusion Eng. Design **83**, 557–572 (2008).
16. Qian, Y.H., d'Humières, D., Lallemand, P.: Lattice BGK models for the Navier–Stokes equation. Europhys. Lett. **17**, 479–484 (1992).
17. Riley, B., Richard, J., Girimaji, S.S.: Assessment of magnetohydrodynamic lattice Boltzmann schemes in turbulence and rectangular jets. Int. J. Mod. Phys. C **19**, 1211–1220 (2008).
18. Succi, S.: The Lattice Boltzmann Equation: For Fluid Dynamics and Beyond. Oxford University Press, Oxford (2001).
19. Vahala, G., Keating, B., Soe, M., Yepez, J., Vahala, L., Carter, J., Ziegeler, S.: MHD turbulence studies using lattice Boltzmann algorithms. Commun. Comput. Phys. **4**, 624–646 (2008).

A-Priori Convergence Analysis of a Discontinuous Galerkin Time-Domain Method to Solve Maxwell's Equations on Hybrid Meshes

C. Durochat and C. Scheid

Abstract We study a multi-element Discontinuous Galerkin Time Domain (DGTD) method for solving the system of unsteady Maxwell equations. This method is formulated on a non-conforming and hybrid mesh combining a structured (orthogonal, large size elements) quadrangulation of the regular zones of the computational domain with an unstructured triangulation for the discretization of the irregularly shaped objects. The main objective is to enhance the flexibility and the efficiency of DGTD methods. Within each element, the electromagnetic field components are approximated by a high order nodal polynomial, using a centered flux for the surface integrals and a second order Leap-Frog scheme for the time integration of the associated semi-discrete equations. We formulate the 3D discretization scheme, present the results of mathematical analysis (L^2 stability and a-priori convergence in 3D). Finally, the 2D numerical performance and convergence is demonstrated.

1 Introduction

Nowadays, a variety of modeling strategies exist for the computer simulation of electromagnetic wave propagation in the time domain. Despite a lot of advances on numerical methods, the FDTD (Finite Difference Time Domain) method [1] is still the prominent modeling approach for realistic time domain computational electromagnetics. The whole computational domain is discretized using a structured

C. Durochat (✉)
Nachos project-team, INRIA Sophia Antipolis – Méditerranée research center 2004 Route des Lucioles, BP 93, 06902 Sophia Antipolis Cedex, France
e-mail: Clement.Durochat@inria.fr

C. Scheid
Jean-Alexandre Dieudonné Mathematics Laboratory, University of Nice – Sophia Antipolis, 06108 Nice Cedex 02, France
e-mail: Claire.SCHEID@unice.fr

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_10,

(Cartesian) grid, which greatly simplifies the discretization process but also represents the main limitation of the method when complicated geometrical objects come into play. Besides, the last 10 years have witnessed an increased interest in so-called DGTD (Discontinuous Galerkin Time Domain) methods. These methods have been developed on quadrangular (2D case) or hexahedral (3D case) meshes [2], as well as on triangular or tetrahedral [3–5] meshes. In all these works on DGTD methods, the first order form of the system of time domain Maxwell equations is considered and within each mesh element the electromagnetic field components are approximated by a arbitrarily high order nodal polynomial. One of the main features of DGTD methods is their enhanced flexibility with regards to the type of meshes they can deal with. Indeed, DGTD methods can easily handle irregular possibly non-conforming meshes. Thus, several attempts have been made to combine time domain methods based on structured meshes with DGTD formulations on unstructured meshes. A low order solution strategy in this direction is presented in [8] in the form of a combination of FDTD and DGTD for 2D problems. The main goal is to accurately model (with DGTD) the geometric details of a curved objects, while maintaining the simplicity and the speed of FDTD for the surrounding space. Furthermore, a high order hybrid strategy has been studied [7] combining a spectral FETD method on quadrangular meshes with a DGTD method on triangular meshes. The coupling of the two methods is achieved by using an upwind numerical flux on the internal boundary defining the interface between the quadrangular and triangular meshes. A different hybridization approach is also recently proposed in [9], which combines the widely adopted Finite Integration Technique (FIT) with a Finite Volume Method (FVM) based either on central or upwind fluxes.

We are concerned here with the possibility of relying on a single discretization scheme, i.e. a DGTD method (denoted by DGTD-$\mathbb{P}_p\mathbb{Q}_k$), and improving the efficiency of the simulation by employing a hybrid structured-unstructured mesh made of orthogonal hexahedral elements (quadrangles in 2D) for the discretization of the regular part of the computational domain and tetrahedral elements (triangles in 2D) for the discretization of the irregularly shaped objects of the propagation scene. The Sect. 2 is devoted to outline the initial and boundary value problem to be solved and the formulation of the DGTD-$\mathbb{P}_p\mathbb{Q}_k$ method in 3D; in Sect. 3 we expose the 3D theoritical L^2 stability and convergence analysis of the scheme, this last leads to a-priori error estimate [5, 6, 11] taking into account the hybrid nature of the mesh; numerical results for a 2D propagation problem are presented in Sect. 4.

2 DGTD-$\mathbb{P}_p\mathbb{Q}_k$ Method on Hybrid Meshes for Maxwell Equations

2.1 The Continuous Maxwell's Problem

Let Ω be an opened, bounded domain of $\mathbb{R}^3$ with boundary Γ. The system of 3D Maxwell equations is given by:

$$\epsilon \partial_t \mathbf{E} - \mathrm{rot}(\mathbf{H}) = -z_0 \sigma \mathbf{E} \ ; \quad \mu \partial_t \mathbf{H} + \mathrm{rot}(\mathbf{E}) = 0, \tag{1}$$

where $\mathbf{E}(\mathbf{x},t) = {}^t\big(E_x(\mathbf{x},t), E_y(\mathbf{x},t), E_z(\mathbf{x},t)\big)$ and $\mathbf{H}(\mathbf{x},t) = {}^t\big(H_x(\mathbf{x},t), H_y(\mathbf{x},t), H_z(\mathbf{x},t)\big)$ respectively denote the electric and magnetic fields (with $\mathbf{x} = {}^t(x_1, x_2, x_3)$); $\epsilon \equiv \epsilon(\mathbf{x})$, $\mu \equiv \mu(\mathbf{x})$ and $\sigma \equiv \sigma(\mathbf{x})$ respectively stand for the electric permittivity, the magnetic permeability and the electric conductivity. Equation (1) have been normalized such that ϵ and μ define relative quantities ($z_0 = \sqrt{\mu_0/\epsilon_0}$ is the vacuum impedance). In this study we only consider PEC boundaries, thus the boundary condition to be applied on Γ reads ($\mathbf{n}$ is the outward normal to Γ) $\mathbf{n} \times \mathbf{E} = 0$. Finally, we assume $\sigma = 0$ and rewrite system (1) under a pseudo-conservative form, where $\mathbf{W} = {}^t(\mathbf{E}, \mathbf{H}) \in \mathbb{R}^6$:

$$Q(\partial_t \mathbf{W}) + \nabla \cdot F(\mathbf{W}) = 0. \tag{2}$$

2.2 DGTD-$\mathbb{P}_p\mathbb{Q}_k$ Space and Time Discretization

The 3D domain Ω is discretized as $\Omega_h = \bigcup_{i=1}^{N} c_i = \mathscr{T}_h \bigcup \mathscr{Q}_h$ where the c_i's are hexahedral ($\in \mathscr{Q}_h$) and tetrahedral ($\in \mathscr{T}_h$) elements. The resulting mesh is hybrid and non-conforming [4], the kind of non-conformity at the interfaces between tetrahedra and hexahedra is described in [10]. Let $\mathbb{P}_p[c_i]$ be the space of polynomial functions with degree at most p in $c_i \in \mathscr{T}_h$, with a local basis $\phi_i = (\varphi_{i1}, \varphi_{i2}, \ldots, \varphi_{id_i})$; and $\mathbb{Q}_k[c_i]$ be the space of polynomial functions with degree at most k with respect to each variable separately in $c_i \in \mathscr{Q}_h$, with a local basis $\theta_i = (\vartheta_{i1}, \ldots, \vartheta_{ib_i})$. The discrete solution vector $\mathbf{W}_h$ is searched for in the approximation space V_h^6 defined by:

$$V_h = \left\{ v_h \in L^2(\Omega) \;\middle|\; \begin{array}{l} \forall c_i \in \mathscr{T}_h,\ v_{h|c_i} \in \mathbb{P}_p[c_i] \\ \forall c_i \in \mathscr{Q}_h,\ v_{h|c_i} \in \mathbb{Q}_k[c_i] \end{array} \right\}.$$

The local degrees of freedom are denoted by $\mathbf{W}_{il} = {}^t(\mathbf{E}_{il}, \mathbf{H}_{il}) \in \mathbb{R}^6$ and $\mathbf{W}_i = {}^t(\mathbf{E}_i, \mathbf{H}_i) \in \mathbb{R}^6$ defines the restriction of the approximate solution to the cell c_i ($\mathbf{W}_{h|c_i}$). When $c_i \in \mathscr{T}_h$, $\mathbf{W}_i$ is defined by $\mathbf{W}_i(\mathbf{x}) = \sum_{l=1}^{d_i} \mathbf{W}_{il}\varphi_{il}(\mathbf{x}) \in \mathbb{P}_p[c_i]$ where d_i is the number of degrees of freedom in the tetrahedron c_i, whereas if $c_i \in \mathscr{Q}_h$, $\mathbf{W}_i$ is defined by $\mathbf{W}_i(\mathbf{x}) = \sum_{l=1}^{b_i} \mathbf{W}_{il}\vartheta_{il}(\mathbf{x}) \in \mathbb{Q}_k[c_i]$ where b_i is the number of degrees of freedom in the hexahedron c_i. Since the approximate fields $\mathbf{E}_h$ and $\mathbf{H}_h$ (i.e. the vector field $\mathbf{W}_h$) are allowed to be completely discontinuous across element boundaries, a specific treatment must be introduced when evaluating such a field at a cell boundary. Let $a_{ij} = c_i \cap c_j$ be the common interface between c_i and c_j and let us denote by $\mathscr{V}_i = \{j \,|\, c_i \cap c_j \neq \varnothing\}$ the set of neighboring cells of c_i. In this study, we choose to use a fully centered numerical flux, i.e. $\forall i, \forall j \in \mathscr{V}_i$ we set $\mathbf{W}_h|_{a_{ij}} = (\mathbf{W}_i|_{a_{ij}} + \mathbf{W}_j|_{a_{ij}})/2$. For the boundary cells i.e. for interfaces

located on the discretization of Γ, we consider that c_j is a fictitious cell and we set $\mathbf{W}_j = {}^t(\mathbf{E}_j,\ \mathbf{H}_j) = {}^t(-\mathbf{E}_i, \mathbf{H}_i)$.

From now on, we consider two cases. Case (A) corresponds to the situation where $c_i \in \mathscr{T}_h$ is a tetrahedron. Then, $\forall j \in \mathscr{V}_i$, a_{ij} is either a boundary interface (i.e. $a_{ij} \in \mathscr{T}_m^i$), or an interface between two neighboring tetrahedra (i.e. $a_{ij} \in \mathscr{T}_d^i$), or a hybrid interface between a tetrahedron and an hexahedron (i.e. $a_{ij} \in \mathscr{H}_d^i$). Let $\overline{\mathbf{E}}_i = {}^t(\mathbf{E}_{i1}, \mathbf{E}_{i2}, \ldots, \mathbf{E}_{id_i})$ and $\overline{\mathbf{H}}_i = {}^t(\mathbf{H}_{i1}, \ldots, \mathbf{H}_{id_i}) \in \mathbb{R}^{3d_i}$, the vectors of local degrees of freedom associated to tetrahedron c_i, while $\widetilde{\mathbf{E}}_j = {}^t(\mathbf{E}_{j1}, \ldots, \mathbf{E}_{jb_j})$ and $\widetilde{\mathbf{H}}_j = {}^t(\mathbf{H}_{j1}, \ldots, \mathbf{H}_{jb_j}) \in \mathbb{R}^{3b_j}$ are the vectors of local degrees of freedom associated to the hexahedron c_j. The weak formulation obtained by dot multiplying (2) by a test function ψ and integrating by parts (not detailed here, cf. [10]) leads to the formulation of a local system of $\mathbf{6d_i}$ semi-discrete equations $\forall c_i \in \mathscr{T}_h$:

$$\begin{cases} 2\mathscr{X}_{\mu,i}\dfrac{d\overline{\mathbf{H}}_i}{dt} - \displaystyle\sum_{k=1}^{3} \mathscr{X}_i^{x_k}\overline{\mathbf{E}}_i - \sum_{a_{ij}\in\mathscr{T}_d^i} \mathscr{X}_{ij}\overline{\mathbf{E}}_j + \sum_{a_{ij}\in\mathscr{T}_m^i} \mathscr{X}_{im}\overline{\mathbf{E}}_i - \sum_{a_{ij}\in\mathscr{H}_d^i} \mathscr{A}_{ij}\widetilde{\mathbf{E}}_j = 0, \\ 2\mathscr{X}_{\epsilon,i}\dfrac{d\overline{\mathbf{E}}_i}{dt} + \displaystyle\sum_{k=1}^{3} \mathscr{X}_i^{x_k}\overline{\mathbf{H}}_i + \sum_{a_{ij}\in\mathscr{T}_d^i} \mathscr{X}_{ij}\overline{\mathbf{H}}_j + \sum_{a_{ij}\in\mathscr{T}_m^i} \mathscr{X}_{im}\overline{\mathbf{H}}_i + \sum_{a_{ij}\in\mathscr{H}_d^i} \mathscr{A}_{ij}\widetilde{\mathbf{H}}_j = 0, \end{cases} \tag{3}$$

where $\mathscr{X}_{\epsilon,i}$ and $\mathscr{X}_{\mu,i}$ are local mass matrices (i.e. involving terms of the form $\int_{\tau_i} {}^t\phi_i\phi_i\, d\mathbf{x}$), the matrix $\mathscr{X}_i^{x_k}$ involves terms of the form $\int_{\tau_i} \left({}^t\phi_i\left(\partial_{x_k}\phi_i\right) - {}^t\left(\partial_{x_k}\phi_i\right)\phi_i\right) d\mathbf{x}$, while $\mathscr{X}_{ij}$ and $\mathscr{X}_{im}$ are matrices associated to boundary integral terms (i.e. involving terms of the form $\int_{a_{ij}} {}^t\phi_i\phi_j\, d\sigma$). All these matrices are of size $3d_i \times 3d_i$ except $\mathscr{A}_{ij}$ whose size is $3d_i \times 3b_j$ (i.e. involving terms of the form $\int_{a_{ij}} {}^t\phi_i\theta_j\, d\sigma$, with $a_{ij} \in \mathscr{H}_d^i$).

In the case (B), $c_i \in \mathscr{Q}_h$ is a hexahedron. Then, $\forall j \in \mathscr{V}_i$, a_{ij} is either a boundary interface, or an interface between two hexahedra, or a hybrid interface ($a_{ij} \in \mathscr{H}_d^i$). We obtain a local system of $\mathbf{6b_i}$ semi-discrete equations $\forall c_i \in \mathscr{Q}_h$, which has the same form than (3) but affecting $\widetilde{\mathbf{E}}_i$ ($\overline{\mathbf{H}}_i$, $\widetilde{\mathbf{E}}_j$ and $\widetilde{\mathbf{H}}_j$, respectively) instead of $\overline{\mathbf{E}}_i$ ($\overline{\mathbf{H}}_i$, $\overline{\mathbf{E}}_j$ and $\overline{\mathbf{H}}_j$, respectively) and $\overline{\mathbf{E}}_j$ ($\overline{\mathbf{H}}_j$, respectively) instead of $\widetilde{\mathbf{E}}_j$ ($\widetilde{\mathbf{H}}_j$, respectively). All matrices (denoted by $\mathscr{W}_{\epsilon,i}$, $\mathscr{W}_{\mu,i}$, etc.) are of size $3b_i \times 3b_i$ and defined analogously to those characterizing the case (A) (but using the local basis θ_i), except the hybrid interface matrix $\mathscr{B}_{ij}$ whose size is $3b_i \times 3d_j$ and defined by ${}^t\mathscr{B}_{ji} = \mathscr{A}_{ij}$.

Finally, time integration of the two systems corresponding to the case (A) and the case (B) relies on a second-order Leap-Frog scheme:

Cas (A):

$$\begin{cases} \overline{\mathbf{H}}_i^{n+\frac{1}{2}} = \overline{\mathbf{H}}_i^{n-\frac{1}{2}} + \dfrac{\Delta t}{2}\left[\mathscr{X}_{\mu,i}\right]^{-1}\mathbf{A}_{\mathbf{E},i}^n, \\ \overline{\mathbf{E}}_i^{n+1} = \overline{\mathbf{E}}_i^n + \dfrac{\Delta t}{2}\left[\mathscr{X}_{\epsilon,i}\right]^{-1}\mathbf{A}_{\mathbf{H},i}^{n+\frac{1}{2}}. \end{cases}$$

Cas (B):

$$\begin{cases} \widetilde{\mathbf{H}}_i^{n+\frac{1}{2}} = \widetilde{\mathbf{H}}_i^{n-\frac{1}{2}} + \dfrac{\Delta t}{2}\left[\mathscr{W}_{\mu,i}\right]^{-1}\mathbf{B}_{\mathbf{E},i}^n, \\ \widetilde{\mathbf{E}}_i^{n+1} = \widetilde{\mathbf{E}}_i^n + \dfrac{\Delta t}{2}\left[\mathscr{W}_{\epsilon,i}\right]^{-1}\mathbf{B}_{\mathbf{H},i}^{n+\frac{1}{2}}. \end{cases}$$

3 Mathematical Analysis

3.1 L^2 Stability Result

For proving the stability of the proposed DGTD-$\mathbb{P}_p\mathbb{Q}_k$ method, we first have to assume following hypothesis, here, *specific to a tetrahedron*:

$$\forall \mathbf{X} \in \left(\mathbb{P}_p[c_i]\right)^3, \quad \|\mathrm{curl}(\mathbf{X})\|_{c_i} \leq \left(\alpha_i^\tau \, p_i \|\mathbf{X}\|_{c_i}\right) / |c_i|,$$
$$\forall \mathbf{X} \in \left(\mathbb{P}_p[c_i]\right)^3, \quad \|\mathbf{X}\|_{a_{ij}}^2 \leq \left(\beta_{ij}^\tau \|\mathbf{n}_{ij}\| \|\mathbf{X}\|_{c_i}^2\right) / |c_i|,$$

where α_i^τ and β_{ij}^τ ($\forall j \in \mathscr{V}_i$) are constant parameters which do not depend on a discretization parameter h, but on the geometry of the finite element and on the interpolation degree. Furthermore, we admit similar hypothesis $\forall \mathbf{X} \in \left(\mathbb{Q}_k[c_i]\right)^3$, where α_i^q and β_{ij}^q defining the constant parameters *specific to a hexahedron*. Besides, $\|.\|_{c_i}$ et $\|.\|_{a_{ij}}$ are L^2 norms, $\mathbf{n}_{ij}$ is the non-unitary normal vector to a_{ij} directed from c_i to c_j with $\|\mathbf{n}_{ij}\| = \int_{a_{ij}} 1 d\sigma$, $|c_i| = \int_{c_i} 1 d\mathbf{x}$, and we also make use $p_i = \sum_{j \in \mathscr{V}_i} \|\mathbf{n}_{ij}\|$. This analysis yields a CFL-like sufficient stability condition, which is:

$$\Delta t = \min(\Delta t_\tau, \Delta t_q) \tag{4}$$

with Δt_τ (specific to triangular part, see [5]) such that $\forall i, \forall j \in \mathscr{V}_i$:

$$\Delta t_\tau \left[2\alpha_i^\tau + \beta_{ij}^\tau \max\left(\sqrt{\epsilon_i/\epsilon_j}, \sqrt{\mu_i/\mu_j}\right)\right] < \left(4|c_i|\sqrt{\epsilon_i \mu_i}\right) / p_i,$$

and Δt_q (specific to quadrangular part) such that $\forall i, \forall j \in \mathscr{V}_i$:

$$\Delta t_q \left[2\alpha_i^q + \beta_{ij}^q \max\left(\sqrt{\epsilon_i/\epsilon_j}, \sqrt{\mu_i/\mu_j}\right)\right] < \left(4|c_i|\sqrt{\epsilon_i \mu_i}\right) / p_i.$$

For the complete proof of this sufficient L^2 stability condition (4), see [10].

3.2 A-Priori Convergence Analysis

To demonstrate the convergence of the DGTD-$\mathbb{P}_p\mathbb{Q}_k$ method, we make several assumptions and introduce some notations. In what follows, h_{c_i} denotes the diameter of the cell (tetrahedral or hexahedral) c_i. We consider a family of unstructed grids $(\mathscr{C}_h)_h$ (hybrid and non-conforming), where h is the mesh parameter of each unstructured grid, defined by $h = \max_{c_i \in \mathscr{C}_h} h_{c_i}$. The meshes $\mathscr{C}_h$ are supposed compatible

with the domain boundary Γ, i.e. the discretized volume $\Omega_h = \bigcup_{c_i \in \mathscr{C}_h}(c_i)$ is equal to Ω. We assume that unstructured grids $\mathscr{C}_h$ are uniformly *shape regular*: there is a constant $\gamma > 0$ such that:

$$\forall h,\ \forall c_i \in \mathscr{C}_h,\ \ h_{c_i} \big/ \eta_{c_i} \ \leq\ \gamma,$$

where η_{c_i} is the diameter of the biggest ball included in the finite element c_i. We also assume the following *inverse assumption*: there is a constant $\zeta > 0$ (independent of h) such that:

$$\forall h,\ \forall c_i \in \mathscr{C}_h,\ \forall j \in \mathscr{V}_i,\ \ h_{c_i} \big/ h_{c_j} \ \leq\ \zeta.$$

We finally shall make the hypothesis that the electromagnetic coefficients ϵ and μ are piecewise constant, we note Ω_j the subdomains of Ω where ϵ and μ are constant.

Next, we introduce the broken Sobolev spaces $PH^{s+1}(\Omega) = \{v \mid \forall j, v_{|\Omega_j} \in H^{s+1}(\Omega_j)\}$ equipped with the norm $\|v\|_{PH^{s+1}(\Omega)} = \left(\sum_j \|v_{|\Omega_j}\|^2_{s+1,\Omega_j}\right)^{1/2}$, where $\|.\|_{s+1,\Omega_j}$ designates the standard H^{s+1}-norm on Ω_j.

From now, let $h_\tau = \max\limits_{c_i \in \mathscr{T}_h} h_{c_i}$ (i.e. the largest diameter of tetrahedral cells), then $h_q = \max\limits_{c_i \in \mathscr{Q}_h} h_{c_i}$ (i.e. the largest diameter of hexahedral cells), and:

$$\xi_h = \max\left\{h_\tau^{\min\{s,p\}}, h_q^{\min\{s,k\}}\right\}.$$

Let $\mathbf{W}_h \in \mathscr{C}^1([0,t_f]; V_h^6)$ and let $\mathbf{W} \in \mathscr{C}^0([0,t_f]; (\ PH^{s+1}(\Omega) \cap H(\mathrm{curl},\Omega)\ )^6)$ for $s \geq 0$ with t_f the final time. Thus, noting C a positive constant independent of ξ_h (and of h), the error $\mathbf{w} = \mathbf{W} - \mathbf{W}_h$ of the semi-discretized scheme satisfies the estimate:

$$\|\mathbf{w}\|_{\mathscr{C}^0([0,t_f],L^2(\Omega))} \ \leq\ C\ \xi_h\ t_f\ \|\mathbf{W}\|_{\mathscr{C}^0([0,t_f],PH^{s+1}(\Omega))}$$

The fully discretized problem may be seen as the discretization in time of a system of ordinary differential equations. There is a local consistency error made by the scheme at each time step. Likewise, since the Leap-Frog scheme is second-order accurate, we found the consistency error altogether of order $\mathscr{O}(\Delta t^2)$. Finally, together with the stability result we thus get an error (under the above assumptions) of order:

$$\mathscr{O}(\Delta t^2) + \mathscr{O}(\xi_h)$$

The complete proof of this theoritical a-priori convergence analysis is in [11].

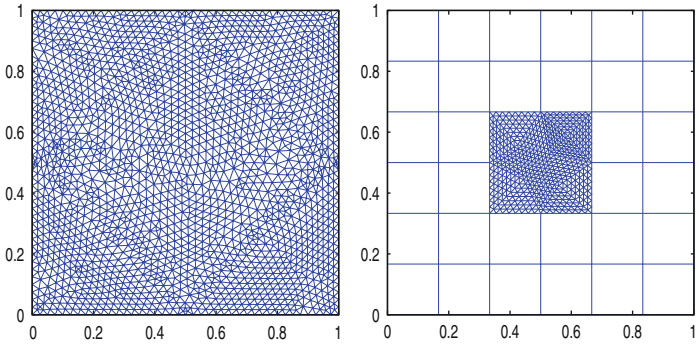


Fig. 1 *Left*: triangular mesh – *Right*: hybrid and non-conforming triangular-quadrangular mesh

Table 1 CPU times, total number of degrees of freedom and L^2-norm of the error at final time

Tri. mesh	CPU time	# dof	Final L^2-error
DGTD-$\mathbb{P}_1$	45 s	11,334	2.33×10^{-2}
DGTD-$\mathbb{P}_2$	206 s	22,668	1.68×10^{-4}
DGTD-$\mathbb{P}_3$	530 s	37,780	7.09×10^{-5}
DGTD-$\mathbb{P}_4$	1,511 s	56,670	2.94×10^{-5}
Hybrid mesh	CPU time	# dof	Final L^2-error
DGTD-$\mathbb{P}_1\mathbb{Q}_4$	12 s	3,872	4.24×10^{-3}
DGTD-$\mathbb{P}_2\mathbb{Q}_3$	45 s	6,656	3.41×10^{-4}
DGTD-$\mathbb{P}_3\mathbb{Q}_4$	160 s	11,040	8.21×10^{-5}
DGTD-$\mathbb{P}_4\mathbb{Q}_4$	346 s	16,160	5.67×10^{-5}

4 2D Numerical Test Case

We consider the case of 2D transverse magnetic waves for which $\mathbf{H} \equiv {}^t\big(H_x, H_y, 0\big)$ and $\mathbf{E} \equiv {}^t(0, 0, E_z)$. The selected test problem is the propagation of an eigenmode in a unitary PEC square cavity. The exact solution is given by:

$$\begin{cases} H_x(x_1, x_2, t) &= -\Big(\pi \big/ \omega\Big) \sin(\pi x_1) \cos(\pi x_2) \sin(\omega t), \\ H_y(x_1, x_2, t) &= \Big(\pi \big/ \omega\Big) \cos(\pi x_1) \sin(\pi x_2) \sin(\omega t), \\ E_z(x_1, x_2, t) &= \sin(\pi x_1) \sin(\pi x_2) \cos(\omega t). \end{cases}$$

where ω is the angular frequency. First we are doing simulations on two meshes (Fig. 1): an unstructured completely triangular mesh and a hybrid triangular-quadrangular (unstructured-structured) and non-conforming mesh (hybrid faces correspond to the non-conforming faces). We test different orders of interpolation and we summarize the results in Table 1. Making comparisons, we observe that the hybridization $\mathbb{P}_1\mathbb{Q}_4$ is more accurate and about four times faster than the interpolation $\mathbb{P}_1$, $\mathbb{P}_2\mathbb{Q}_3$ is here more accurate by about a factor 70 compared to $\mathbb{P}_1$ with the same CPU time. $\mathbb{P}_3\mathbb{Q}_4$ ($\mathbb{P}_4\mathbb{Q}_4$, respectively) is also more accurate and faster than $\mathbb{P}_2$ ($\mathbb{P}_3$, respectively).

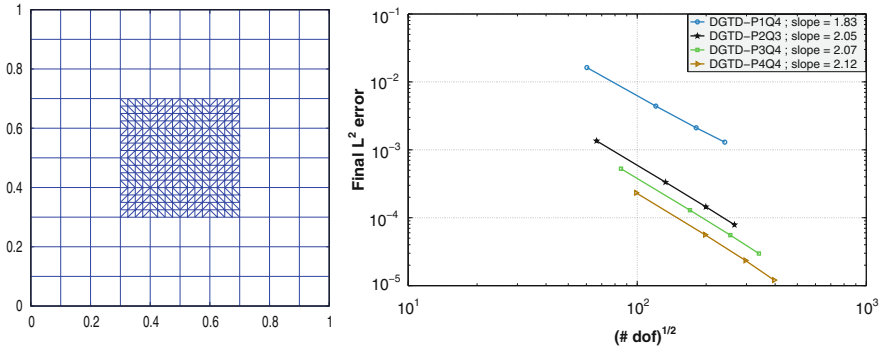


Fig. 2 *Left*: hybrid and non-conform mesh based on 11×11 points (same form for the three others meshes) – *Right*: final L^2-error according to square root of the total number of degrees of freedom

Now, we study the h-wise numerical convergence for these interesting hybridizations. We perform tests on four meshes based on these resolutions: 11×11 (Fig. 2), 21×21, 31×31 and 41×41 points. We measure the L^2-norm of the error at final time for these different resolutions and we calculate the corresponding slopes (Fig. 2). We clearly observe that the numerical convergence is validated for the four hybridizations. For each of these, the order of convergence (the slope on plot) 2 is achieved (except slightly less for $\mathbb{P}_1\mathbb{Q}_4$ because of the $\mathbb{P}_1$ interpolation on the triangular part). Finally, we note that for each hybridization $\mathbb{P}_p\mathbb{Q}_k$, the global time step used is the minimum between the time steps of the triangular and quadrangular parts, thus validating the sufficient condition resulting from the stability analysis.

5 Conclusion and Future Research Tracks

In this paper, we have presented the results of an investigation of a multi-element DGTD-$\mathbb{P}_p\mathbb{Q}_k$ method on hybrid and non-conforming meshes solving the time-domain Maxwell's equations. We were essentially concerned here with the a-priori convergence analysis for the 3D case, the proof of this analysis will appear soon in [11], and we have studied a 2D numerical test problem giving promising results for different hybridizations. Future works will aim at a detailed assessment of the method for more realistic problems, in 2D first, and we will extend the method to numerical problems in 3D. Also, the computational efficiency can be further improved while minimizing the dispersion error thanks to the use of orthogonal basis functions for $\mathbb{Q}_k$ interpolation on hexahedra and a local time stepping strategy.

Acknowledgements The authors gratefully acknowledge support from Région Ile-de-France in the framework of the MIEL3D-MESHER project of the System@tic Paris-Région cluster.

References

1. Taflove, A., Hagness, S.C.: Computational electrodynamics: the finite-difference time-domain method - 3rd ed. Artech House Publishers (2005)
2. Cohen, G., Ferrieres, X., Pernet, S.: A spatial high order hexahedral discontinuous Galerkin method to solve Maxwell's equations in time-domain. J. Comput. Phys. **217**(2), 340–363 (2006)
3. Hesthaven, J.S., Warburton, T.: Nodal high-order methods on unstructured grids. I. Time-domain solution of Maxwell's equations. J. Comput. Phys. **181**(1), 186–221 (2002)
4. Fahs, H., Lanteri, S.: A high-order non-conforming discontinuous Galerkin method for time-domain electromagnetics. J. Comput. Appl. Math. **234**, 1088–1096 (2010) doi:10.1016/j.cam.2009.05.015
5. Fezoui, L., Lanteri, S., Lohrengel, S., Piperno, S.: Convergence and stability of a discontinuous Galerkin time-domain method for the 3D heterogeneous Maxwell equations on unstructured meshes. ESAIM: Math. Model. Numer. Anal. **39**(6), 1149–1176 (2005)
6. Pernet, S., Ferrieres, X.: Hp a-priori error estimates for a non-dissipative spectral discontinuous Galerkin method to solve the Maxwell equations in the time domain. Math. Comput. **76**, 1801–1832 (2007)
7. Davies, R.W., Morgan, K., Hassan, O.: A high order hybrid finite element method applied to the solution of electromagnetic wave scattering problems in the time domain. Comput. Mech. **44**, 321–331 (2009)
8. Garcia, S.G., Pantoja, M.F., de Jong van Coevorden, C.M., Bretones, A.R., Martin, R.G.: A new hybrid DGTD/FDTD method in 2-D. IEEE Microw. Wireless Compon. Lett. **18**(12), 764–766 (2008)
9. Schnepp, S., Gjonaj, E., Weiland, T.: A hybrid finite integration. J. Comput. Phys. **229**(11), 4075–4096 (2010)
10. Durochat, C., Lanteri, S.: Discontinuous Galerkin method on hybrid meshes triangular / quadrangular for the numerical resolution of the time domain Maxwell's equations. INRIA Research report no. 7253 (2010)
11. Durochat, C., Scheid, C.: A priori convergence analysis of a Non-Conforming Discontinuous Galerkin Time-Domain method on hybrid meshes for the Maxwell equations. INRIA Research report no. 7933 (2012)

Stabilization of a Degenerate Minimization Problem with the Single-Layer Potential

S. Ferraz-Leite

Abstract We consider the reduced model for thin-film devices in stationary micromagnetics proposed in DeSimone et al. (R Soc Lond Proc A 457(2016):2983–2991, 2001). In the case of *soft* material, one of the energy contributions is negligible, and the problem becomes degenerate. The analysis and the numerical scheme recently developed in Ferraz-Leite et al. (Numer Math, Accepted for publication, 2012) are not satisfactory in this case. In the present work, we overcome the degeneracy and extend the numerical scheme by introducing a stabilizing energy term. Convergence of the method is established and a numerical experiment concludes the paper.

1 Model Problem and Introduction

We consider the model that was proposed in [3, 4] for the simulation of thin ferromagnetic films: Let $\Omega = \omega \times (0, t)$ be a thin ferromagnetic sample. The surface $\omega \subseteq \mathbb{R}^2$ is a Lipschitz domain with $\operatorname{diam}(\omega) \sim 1$. We model the sample as this screen and neglect the thickness $t \ll 1$.

We are interested in the effective behavior of the ferromagnetic material when exposed to an in-plane exterior field $\mathbf{h}_{\text{ext}}$. For simplicity we assume $\mathbf{h}_{\text{ext}} \in \mathbb{R}^2$ to be constant. The magnetization $\mathbf{m} : \omega \to \mathbb{R}^2$ is an in-plane vector field. In the full 3-dimensional model due to Landau and Lifschits [8], the magnetization is of constant length $|\mathbf{m}(x)| = 1$. In the reduced thin-film model, however, this constraint relaxes to the convex admissibility condition $|\mathbf{m}(x)| \leq 1$ almost everywhere in ω.

Let V denote the single-layer operator on ω corresponding to the Laplacian in 3D, i.e.

$$(V\varphi)(x) := \frac{1}{4\pi} \int_\omega \frac{\varphi(y)}{|x - y|}\, dy, \quad x \in \omega. \tag{1}$$

S. Ferraz-Leite (✉)
Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, 04103 Leipzig, Germany
e-mail: samuel.ferraz-leite@mis.mpg.de

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_11,

We recall that $V \in L(\widetilde{H}^{-1/2}(\omega); H^{1/2}(\omega))$ is elliptic and the induced norm $\|\varphi\|_V^2 := \langle \varphi, V\varphi \rangle_{\widetilde{H}^{-1/2}(\omega)\times H^{1/2}(\omega)}$ is an equivalent norm on the negative fractional order Sobolev space $\widetilde{H}^{-1/2}(\omega) := (H^{1/2}(\omega))^*$, see e.g. [10, 11].

The steady states of the magnetization $\mathbf{m}$ in presence of the applied field $\mathbf{h}_{\text{ext}}$ are minimizers of the quadratic energy functional

$$e(\mathbf{m}) = \frac{1}{2}\|\nabla \cdot \mathbf{m}\|_V^2 + \frac{q}{2}\|\mathbf{m}_2\|_{L^2(\omega)}^2 - (\mathbf{h}_{\text{ext}}, \mathbf{m})_{L^2(\omega)^2}. \tag{2}$$

The first term is the so-called stray-field energy. In the thin-film model, the magnetostatic potential $u : \mathbb{R}^3 \to \mathbb{R}$ is given as single-layer potential of the negative divergence of $\mathbf{m}$, i.e.

$$u(x) = -\frac{1}{4\pi}\int_\omega \frac{\nabla \cdot \mathbf{m}(y)}{|x-y|}\, dy, \quad x \in \mathbb{R}^3 \setminus \overline{\omega}. \tag{3}$$

This representation leads to the energy contribution $\|\nabla \cdot \mathbf{m}\|_V^2$. The second term models a crystalline anisotropy of the material. Here, we consider the uniaxial case where the magnetization is favored to align with the first in-plane axis, and the energy contribution is scaled with a material-parameter $q \geq 0$. Finally, the linear term favors alignment of $\mathbf{m}$ along the applied field $\mathbf{h}_{\text{ext}}$.

In [6, 7] the authors establish an appropriate Hilbert space setting and analyze well-posedness of the model problem. The energy space for the magnetization is $\mathscr{H} := \{\mathbf{m} \in L^2(\omega)^2 \mid \nabla\cdot\mathbf{m} \in \widetilde{H}^{-1/2}(\omega), \mathbf{m}\cdot\mathbf{n} = 0 \text{ on } \partial\omega\}$, where $\mathbf{n}$ denotes the outer normal vector of $\omega \subseteq \mathbb{R}^2$. The space is naturally equipped with the norm $\|\mathbf{m}\|_{\mathscr{H}} := (\|\mathbf{m}\|_{L^2}^2 + \|\nabla \cdot \mathbf{m}\|_{\widetilde{H}^{-1/2}}^2)^{1/2}$. The admissible set is $\mathscr{A} := \{\mathbf{m} \in \mathscr{H} \mid |\mathbf{m}| \leq 1\}$.

Theorem 1 ([7, Theorem 11]). *The energy $e(\mathbf{m})$ of (2) admits a minimizer $\mathbf{m}^*$ in $\mathscr{A}$. The minimizer depends continuously on the data $\mathbf{h}_{\text{ext}}$ with respect to the energy semi-norm $|||\mathbf{m}||| := (\|\nabla \cdot \mathbf{m}\|_V^2 + q\|\mathbf{m}_2\|_{L^2(\omega)}^2)^{1/2}$. For $q > 0$, the quantity $|||\mathbf{m}|||$ is a (not equivalent) norm on $\mathscr{H}$, and the minimizer $\mathbf{m}^*$ is uniquely determined.*

In [5, 7], a penalty method in the spirit of [1] was proposed to solve the model problem. Although the analysis formally covers the entire parameter regime $q \geq 0$, it is tailored for the case $q > 0$. From a mathematical point of view the model is better justified for *soft* ferromagnetic films [3], i.e. $q = 0$. In this case, however, $|||\mathbf{m}|||$ is only a semi-norm but not a norm; minimizers $\mathbf{m}^*$ are not uniquely determined; the problem becomes degenerate. In the present work, we closely analyze this specific case $q = 0$. First, we establish uniqueness of a minimum-norm solution $\mathbf{m}^{**}$. Then, for some $\delta > 0$, we introduce the stabilized energy

$$e^\delta(\mathbf{m}) := \frac{1}{2}\|\nabla \cdot \mathbf{m}\|_V^2 + \frac{\delta}{2}\|\mathbf{m}\|_{L^2(\omega)}^2 - (\mathbf{h}_{\text{ext}}, \mathbf{m})_{L^2(\omega)}. \tag{4}$$

This stabilized energy admits a unique minimizer $\mathbf{m}^\delta$ in $\mathscr{A}$, and we prove convergence

$$\mathbf{m}^\delta \rightharpoonup \mathbf{m}^{**} \quad \text{as} \quad \delta \to 0. \tag{5}$$

2 Degeneracy and the Penalty Method

In this Section we recall in short the penalty scheme from [7] for the approximation of solutions $\mathbf{m}^*$. We point out the difficulties and restrictions in the analysis that arise for $q = 0$. In this case the energy functional (2) reads

$$e(\mathbf{m}) = \frac{1}{2}\|\nabla \cdot \mathbf{m}\|_V^2 - (\mathbf{h}_{\text{ext}}, \mathbf{m})_{L^2(\omega)^2}. \tag{6}$$

As a consequence of the assumption that $\mathbf{h}_{\text{ext}}$ is constant, the constraint $\mathbf{m} \cdot \mathbf{n} = 0$ on $\partial\omega$ and integration by parts yield

$$\int_\omega \mathbf{h}_{\text{ext}} \cdot \mathbf{m}\, dx = -\int_\omega (\mathbf{h}_{\text{ext}} \cdot x)\nabla \cdot \mathbf{m}\, dx, \tag{7}$$

and hence the energy $e(\mathbf{m})$ can be expressed only in terms of $\nabla \cdot \mathbf{m}$ by

$$e(\mathbf{m}) = e(-\nabla \cdot \mathbf{m}) = \frac{1}{2}\|\nabla \cdot \mathbf{m}\|_V^2 + (\mathbf{h}_{\text{ext}} \cdot x, -\nabla \cdot \mathbf{m})_{L^2(\omega)}. \tag{8}$$

We use the notation $\nabla^\perp \psi = (-\partial_2 \psi, \partial_1 \psi)^T$ and observe $\nabla \cdot (\nabla^\perp \psi) = 0$. Let $\mathbf{m}^*$ be a minimizer of the energy (6). Then, any stream function $\psi \in H^1(\omega)$ such that $(\mathbf{m}^* + \nabla^\perp \psi) \in \mathscr{A}$ yields another admissible minimizer $\mathbf{m}^* + \nabla^\perp \psi$. In general we cannot expect the solution $\mathbf{m}^*$ to be unique. But since $\|\varphi\|_V$ is an equivalent norm on $\widetilde{H}^{-1/2}(\omega)$, at least the divergence of a minimizer $\nabla \cdot \mathbf{m}^*$ is unique.

We stress that the rough geometry of $\mathscr{A}$ does not allow a straight forward use of projection based schemes. In order to compute a numerical approximation to some minimizer $\mathbf{m}^*$, we first propose a penalty scheme on the continuous level and then discretize the resulting problem conformingly. We define the positive part function

$$(u)_+(x) := \begin{cases} u(x), & \text{if } u(x) \geq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Given $\varepsilon > 0$, we seek a minimizer $\mathbf{m}^\varepsilon \in \mathscr{H}$ of the penalized energy

$$e_\varepsilon(\mathbf{m}) := e(\mathbf{m}) + \frac{1}{2\varepsilon}\|(|\mathbf{m}| - 1)_+\|_{L^2(\Omega)}^2. \tag{10}$$

The penalized energy functional (10) is convex and coercive in $\mathscr{H}$ and, hence, admits a (non-unique) minimizer $\mathbf{m}^\varepsilon \in \mathscr{H}$.

To solve the unconstrained non-linear penalized problem, we discretize the energy space $\mathscr{H}$ conformingly: Let $\mathscr{T}_h$ denote a regular triangulation of ω in the sense of Ciarlet [2] with mesh-size $h > 0$. We denote by $RT^0(\mathscr{T}_h)$ the space of lowest-order Raviart-Thomas finite element functions [9]. This is a natural conforming discretization of $\mathscr{H}$ [6], and since $RT^0(\mathscr{T}_h) \subseteq \mathscr{H}$ is a closed subspace, we immediately conclude existence of a (non-unique) minimizer

$$\mathbf{m}_h^\varepsilon = \operatorname{argmin}_{RT^0(\mathscr{T}_h)} e_\varepsilon(\mathbf{m}_h). \tag{11}$$

In [7] the authors prove that this approach—penalization and conforming discretization—yields an unconditionally convergent scheme, but the notion of convergence for $q = 0$ is quite unsatisfying.

Theorem 2 ([7, Theorem 11]). *Let $(h_n)_{n\in\mathbb{N}}$ and $(\varepsilon_n)_{n\in\mathbb{N}}$ be arbitrary positive zero sequences, i.e., $h_n, \varepsilon_n > 0$ with $\lim_{n\to\infty} h_n = \lim_{n\to\infty} \varepsilon_n = 0$. A sequence of minimizers $\mathbf{m}_n \in RT^0(\mathscr{T}_{h_n})$ of the penalized energy $e_{\varepsilon_n}(\cdot)$ from (10) satisfies convergence in the following sense: Any subsequence $(\mathbf{m}_k)_{k\in\mathbb{N}} \subseteq (\mathbf{m}_n)_{n\in\mathbb{N}}$ contains a convergent subsequence $(\mathbf{m}_\ell)_{\ell\in\mathbb{N}} \subseteq (\mathbf{m}_k)_{k\in\mathbb{N}}$ whose limit is a minimizer $\mathbf{m}^* \in \mathscr{A}$ of the energy $e(\cdot)$ from (2). Convergence holds with respect to the weak topology of $\mathscr{H}$ and the topology induced by the semi-norm $|||\mathbf{m}||| := (\|\nabla\cdot\mathbf{m}\|_V^2 + q\|\mathbf{m}_2\|_{L^2(\omega)}^2)^{1/2}$ i.e.*

$$\mathbf{m}_\ell \rightharpoonup \mathbf{m}^* \qquad \text{and} \qquad |||\mathbf{m}_\ell - \mathbf{m}^*||| \to 0. \tag{12}$$

Moreover, the entire sequence of energies converges:

$$e(\mathbf{m}_n) \to e(\mathbf{m}^*) \quad \text{and} \quad e_{\varepsilon_n}(\mathbf{m}_n) \to e(\mathbf{m}^*). \tag{13}$$

For $q > 0$, the minimizer $\mathbf{m}^*$ is uniquely determined and we have convergence of the full sequence, i.e. (12) holds with $\mathbf{m}_\ell$ replaced by $\mathbf{m}_n$.

Remark 1. The degeneracy in the case $q = 0$ has the following consequences:

- To compute a minimizer $\mathbf{m}_n \in RT^0(\mathscr{T}_{h_n})$ we use a damped Newton-method. The Hessian of the penalized energy reads

$$\operatorname{Hess}(e(\mathbf{m}_n))(\mathbf{v})(\mathbf{w}) = \langle \nabla\cdot\mathbf{v}, V(\nabla\cdot\mathbf{w})\rangle + (f_\varepsilon(\mathbf{m}_n)\mathbf{v}, \mathbf{w})_{L^2(\omega)} \tag{14}$$

  with some non-linearity $f_\varepsilon(\mathbf{m}_n)$ that corresponds to the penalty term. This is in general not a positive definite form, but merely positive *semi-definite*.
- Convergence is mathematically only ensured for some subsequence $\mathbf{m}_\ell$. The numerical approximations could oscillate between different minimizers $\mathbf{m}^*$ of the continuous and constrained problem. In experiments, we observe that different initial values for the Newton algorithm lead to different minimizers.

3 A Stabilized Approximation

In this Section we define the minimum norm minimizer and propose a stabilized scheme to compute it. The minimum norm solution is of special physical relevance. It allows reconstruction of certain microstructural phenomena from the effective magnetization computed by the reduced model, cf. [4].

3.1 Minimum Norm Solution and Stabilized Energy Functional

As already mentioned, the quantity $\sigma^* = -\nabla \cdot \mathbf{m}^*$ is uniquely determined from the model problem (6). We minimize the L^2-norm of $\mathbf{m} \in \mathscr{H}$ under the side constraints $\mathbf{m} \in \mathscr{A}$ and $-\nabla \cdot \mathbf{m} = \sigma^*$ and define the solution $\mathbf{m}^{**}$ as minimum norm minimizer.

Proposition 1. *Let $\sigma \in \widetilde{H}^{-1/2}(\omega)$ be given. Then there is a uniquely determined $\mathbf{m}^{**} \in \mathscr{A}_\sigma := \{\mathbf{m} \in \mathscr{A} \mid -\nabla \cdot \mathbf{m} = \sigma\}$ such that $\|\mathbf{m}^{**}\|_{L^2(\omega)} \leq \|\mathbf{m}\|_{L^2(\omega)}$ for $\mathbf{m} \in \mathscr{A}_\sigma$.*

Proof. We recall that $\mathscr{A}$ is a closed and convex subset of $\mathscr{H}$, cf. [7]. The set of functions $\mathbf{m} \in \mathscr{H}$ that satisfy $-\nabla \cdot \mathbf{m} = \sigma$ can be written as $\mathscr{H}_\sigma := \{\mathbf{m} \in \mathscr{H} \mid \mathbf{m} = \nabla\phi + \mathbf{w}$ with $\nabla \cdot \mathbf{w} = 0\}$. The function $\phi \in H^1(\omega)$ is given as the (up to constants uniquely determined) solution to the Neumann problem

$$-\Delta\phi = \sigma \in \widetilde{H}^{-1}(\omega) \supseteq \widetilde{H}^{-1/2}(\omega), \tag{15}$$

$$\frac{\partial\phi}{\partial\mathbf{n}} = 0 \in \widetilde{H}^{-1/2}(\partial\omega). \tag{16}$$

The admissibility set $\mathscr{A}_\sigma := \mathscr{A} \cap \mathscr{H}_\sigma$ is closed and convex with respect to the norm topology of $\mathscr{H}$. The functional $\mathbf{m} \mapsto \|\mathbf{m}\|_{L^2}^2$ is obviously strictly convex, continuous on $\mathscr{H}$, and coercive on $\mathscr{A}_\sigma$. □

From a numerical point of view this approach is not suitable. First, it doesn't provide uniqueness of $\mathbf{m}^*$ in the first step for the computation of σ^*, and second, the side constraint $-\nabla \cdot \mathbf{m} = \sigma^*$ seems numerically inconvenient. Instead, we introduce the stabilized energy functional

$$e^\delta(\mathbf{m}) = \frac{1}{2}\|\nabla \cdot \mathbf{m}\|_V^2 + \frac{\delta}{2}\|\mathbf{m}\|_{L^2(\omega)}^2 - (\mathbf{h}_{\text{ext}}, \mathbf{m})_{L^2(\omega)}. \tag{17}$$

This depends on a (small) parameter $\delta > 0$ and we seek to compute minimizers $\mathbf{m} \in \mathscr{A}$ of the energy (17).

Proposition 2. *There is a unique minimizer $\mathbf{m}^\delta \in \mathscr{A}$ of the energy (17).*

Proof. We define the δ-norm

$$|||\mathbf{m}|||_\delta^2 = \|\nabla\cdot\mathbf{m}\|_V^2 + \delta\|\mathbf{m}\|_{L^2(\omega)}^2 \tag{18}$$

and observe that it is an equivalent norm on $\mathscr{H}$. We seek to minimize

$$e^\delta(\mathbf{m}) = \frac{1}{2}|||\mathbf{m}|||_\delta^2 - F(\mathbf{m}) \tag{19}$$

where $F \in L(\mathscr{H},\mathbb{R})$ is defined through $F(\mathbf{m}) = (\mathbf{h}_{\text{ext}},\mathbf{m})_{L^2(\omega)}$. Obviously this is a strictly convex, continuous, and coercive minimization problem. Since $\mathscr{A}$ is a closed and convex subset of $\mathscr{H}$, the problem has a unique solution $\mathbf{m}^\delta \in \mathscr{A}$. □

3.2 Convergence

We now establish convergence of $\mathbf{m}^\delta$ to the minimum norm solution $\mathbf{m}^{**}$ as $\delta \to 0$. The proof is organized in two steps: First, energy bounds provide weak convergence to some minimizer $\mathbf{m}^*$. From that, we obtain convergence of the quantity $\sigma^\delta = -\nabla\cdot\mathbf{m}^\delta$ to the uniquely determined $\sigma^* = -\nabla\cdot\mathbf{m}^*$ in the $\widetilde{H}^{-1/2}(\omega)$-norm. Second, the stabilized energy functional (17) provides the lower bound $\|\mathbf{m}^\delta\|_{L^2(\omega)} \le \|\mathbf{m}^{**}\|_{L^2(\omega)}$. Together, we get that each accumulation-point $\widetilde{\mathbf{m}}$ of the sequence $\mathbf{m}^\delta$ is a minimizer of the energy (6) such that $\|\widetilde{\mathbf{m}}\|_{L^2(\omega)} \le \|\mathbf{m}^{**}\|_{L^2(\omega)}$, which proves $\widetilde{\mathbf{m}} = \mathbf{m}^{**}$.

Before stating the convergence of the proposed scheme, we recall two simple Lemmas. The proofs are left to the reader.

Lemma 1. *Let $(x_n)_{n\in\mathbb{N}} \subseteq X$ be a sequence and $x \in X$ some element of the metric space (X,d). Assume that each subsequence $(x_k)_{k\in\mathbb{N}} \subseteq (x_n)_{n\in\mathbb{N}}$ has a subsequence $(x_\ell)_{\ell\in\mathbb{N}} \subseteq (x_k)_{k\in\mathbb{N}}$ such that $\lim_{\ell\to\infty} x_\ell = x$. Then we already have $\lim_{n\to\infty} x_n = x$.* □

Lemma 2. *Let $(u_n)_{n\in\mathbb{N}} \subseteq H$ be a sequence in a Hilbert space $(H,\|\cdot\|_H)$. Let $(\cdot,\cdot)_H$ denote a continuous semi-scalar product with induced semi-norm $|\cdot|_H = (\cdot,\cdot)_H^{1/2}$. Assume that the sequence has a weak limit $u_n \rightharpoonup u$ and that the semi-norm converges $\lim_{n\in\mathbb{N}} |u_n|_H = |u|_H$. Then we already have $\lim_{n\to\infty} |u_n - u|_H = 0$.* □

Theorem 3. *Let $(\delta_n)_{n\in\mathbb{N}}$ be some positive zero sequence, i.e. $\delta_n > 0$ and $\lim_{n\to\infty}\delta_n = 0$. Then, the sequence $(\mathbf{m}^{\delta_n})_{n\in\mathbb{N}} \subseteq \mathscr{A}$ of minimizers of the stabilized energy (17) satisfies*

$$\|\nabla\cdot(\mathbf{m}^{\delta_n} - \mathbf{m}^{**})\|_{\widetilde{H}^{-1/2}(\omega)} \to 0, \qquad \lim_{n\to\infty} \mathbf{m}^{\delta_n} = \mathbf{m}^{**}, \tag{20}$$

where the second statement holds with respect to the weak topology of $\mathscr{H}$.

Proof. We may assume without loss of generality $\delta_n \le 1$. From $e^\delta(\mathbf{m}) = e(\mathbf{m}) + \frac{\delta}{2}\|\mathbf{m}\|_{L^2(\omega)}^2$ and the fact that $\mathbf{m}^{\delta_n} \in \mathscr{A}$ is the minimizer of e^{δ_n}, we conclude

$$e(\mathbf{m}^{\delta_n}) \le e^{\delta_n}(\mathbf{m}^{\delta_n}) \le e^{\delta_n}(\mathbf{m}^{**}) \le e^1(\mathbf{m}^{**}) = K. \tag{21}$$

From coercivity of $e(\cdot)$ we obtain boundedness of $(\mathbf{m}^{\delta_n})_{n\in\mathbb{N}}$ in $\mathscr{H}$. Therefore, each subsequence $(\mathbf{m}^{\delta_k})_{k\in\mathbb{N}} \subseteq (\mathbf{m}^{\delta_n})_{n\in\mathbb{N}}$ has a weakly convergent subsequence $(\mathbf{m}^{\delta_\ell})_{\ell\in\mathbb{N}} \subseteq (\mathbf{m}^{\delta_k})_{k\in\mathbb{N}}$. We choose one such subsequence $(\mathbf{m}^{\delta_\ell})_{\ell\in\mathbb{N}}$ and denote its weak limit by $\widetilde{\mathbf{m}} \leftharpoonup \mathbf{m}^{\delta_\ell}$. Trivially, $e(\mathbf{m}^{**}) \le e(\widetilde{\mathbf{m}})$. Weak lower semicontinuity of $e(\cdot)$ yields the converse inequality

$$e(\widetilde{\mathbf{m}}) \le \liminf_{\ell\in\mathbb{N}} e(\mathbf{m}^{\delta_\ell}) \le \liminf_{\ell\in\mathbb{N}} e^{\delta_\ell}(\mathbf{m}^{\delta_\ell}) \le \liminf_{\ell\in\mathbb{N}} e^{\delta_\ell}(\mathbf{m}^{**}) = e(\mathbf{m}^{**}), \tag{22}$$

i.e. the weak limit $\widetilde{\mathbf{m}}$ is a minimizer of the energy $e(\cdot)$.

From the uniqueness of $\sigma^* = -\nabla\cdot\mathbf{m}^{**}$, we conclude $-\nabla\cdot\widetilde{\mathbf{m}} = \sigma^*$. Recall that

$$(\mathbf{m},\mathbf{w})_\sigma := \langle \nabla\cdot\mathbf{m}, V(\nabla\cdot\mathbf{w})\rangle_{\widetilde{H}^{-1/2}(\omega)\times H^{1/2}(\omega)}, \tag{23}$$

defines a continuous semi-scalar product on the energy space $\mathscr{H}$. We denote the induced continuous semi-norm by $|\cdot|_\sigma$. The energy $e(\cdot)$ can be written as

$$e(\mathbf{m}) = \frac{1}{2}|\mathbf{m}|_\sigma - (\mathbf{h}_{\text{ext}},\mathbf{m})_{L^2(\omega)}. \tag{24}$$

The mapping $\mathbf{m} \mapsto (\mathbf{h}_{\text{ext}},\mathbf{m})_{L^2(\omega)}$ is a linear and continuous functional on $\mathscr{H}$. From $e(\widetilde{\mathbf{m}}) \le \liminf_{\ell\in\mathbb{N}} e(\mathbf{m}^{\delta_\ell})$ and $e^{\delta_\ell}(\mathbf{m}^{\delta_\ell}) \le e^{\delta_\ell}(\mathbf{m}^{**}) \to e(\mathbf{m}^{**})$, we conclude convergence of the energy $\lim_{\ell\to\infty} e(\mathbf{m}^{\delta_\ell}) = e(\mathbf{m}^{**})$. From the weak convergence we have $(\mathbf{h}_{\text{ext}},\mathbf{m}^{\delta_\ell})_{L^2(\omega)} \to (\mathbf{h}_{\text{ext}},\widetilde{\mathbf{m}})_{L^2(\omega)}$, and we finally obtain convergence of the semi-norm $\lim_{\ell\to\infty}|\mathbf{m}^{\delta_\ell}|_\sigma = |\mathbf{m}^{**}|_\sigma$. With Lemma 2, we conclude $\nabla\cdot\mathbf{m}^{\delta_n} \to \nabla\cdot\mathbf{m}^{**} \in \widetilde{H}^{-1/2}(\omega)$. Consider the sequence $(\sigma_n := -\nabla\cdot\mathbf{m}^{\delta_n})_{n\in\mathbb{N}}$. According to Lemma 1, we have proven $\lim_{n\to\infty}\sigma_n = \sigma^* \in \widetilde{H}^{-1/2}(\omega)$.

Next, we show that the stabilized energy functional in fact approximates the minimum norm solution from the set of available choices. We use the fact that $\mathbf{m}^{**}$ and $\mathbf{m}^{\delta_n}$ are minimizers of the corresponding energy functionals to see

$$e(\mathbf{m}^{**}) \le e(\mathbf{m}^{\delta_n}) \quad\text{and}\quad e(\mathbf{m}^{\delta_n}) + \frac{\delta}{2}\|\mathbf{m}^{\delta_n}\|^2_{L^2(\omega)} \le e(\mathbf{m}^{**}) + \tfrac{\delta}{2}\|\mathbf{m}^{**}\|^2_{L^2(\omega)}. \tag{25}$$

This means that

$$\frac{\delta}{2}\left(\|\mathbf{m}^{\delta_n}\|^2_{L^2(\omega)} - \|\mathbf{m}^{**}\|^2_{L^2(\omega)}\right) \le e(\mathbf{m}^{**}) - e(\mathbf{m}^{\delta_n}) \le 0, \tag{26}$$

and hence $\|\mathbf{m}^{\delta_n}\|_{L^2(\omega)} \le \|\mathbf{m}^{**}\|_{L^2(\omega)}$. Now consider again the weakly convergent subsequence $(\mathbf{m}^{\delta_\ell})_{\ell\in\mathbb{N}}$ whose weak limit $\widetilde{\mathbf{m}}$, as we have seen before, is a minimizer. The L^2-norm is convex and continuous on $\mathscr{H}$. It is, thus, weakly lower semicontinuous, and we easily conclude $\|\widetilde{\mathbf{m}}\|_{L^2} \le \liminf_\ell \|\mathbf{m}^{\delta_\ell}\|_{L^2} \le \|\mathbf{m}^{**}\|_{L^2}$.

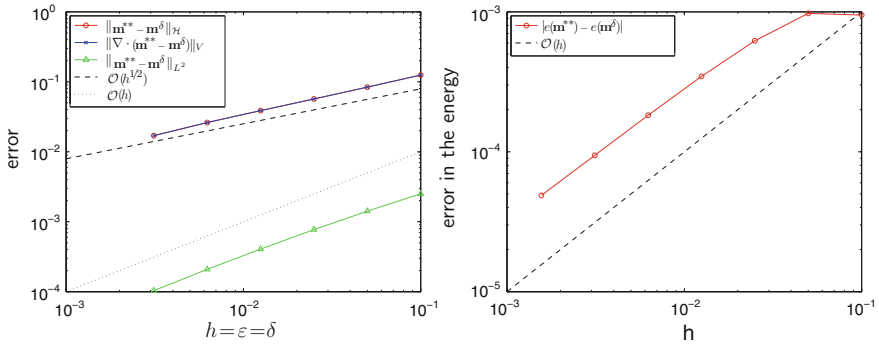


Fig. 1 *Left*: error $(\mathbf{m}^{**}-\mathbf{m}^{\delta})$ measured in the $\mathscr{H}$-, V-, and L^2-norm. *Right*: error in the energy $|e(\mathbf{m}^{**})-e(\mathbf{m}^{\delta})|$. All quantities are plotted over the mesh-size h. The parameters are $\varepsilon=\delta=h$

Since we know $\nabla\cdot\widetilde{\mathbf{m}}=\nabla\cdot\mathbf{m}^{**}$ and according to Proposition 1, this means $\widetilde{\mathbf{m}}=\mathbf{m}^{**}$. Finally, Lemma 1 yields $\lim_{n\to\infty}\mathbf{m}^{\delta_n}=\mathbf{m}^{**}\in\mathscr{H}$ in the weak topology. □

Remark 2. We only established convergence of the sequence of continuous minimizers $\mathbf{m}^{\delta}$. Similar ideas as in [7] can be used to prove that a sequence $\mathbf{m}_n=\mathbf{m}_{h_n}^{\varepsilon_n,\delta_n}$ of discrete minimizers of the stabilized and penalized problem converges

$$\mathbf{m}_n\to\mathbf{m}^{**}\quad\text{as}\quad(h_n,\varepsilon_n,\delta_n)\to(0,0,0)\tag{27}$$

without additional assumptions on the zero sequences $h_n,\varepsilon_n,\delta_n$. A close analysis of the Euler-Lagrange equations as in [5]—a topic that would exceed the scope of this paper—leads to the heuristic choice of $h\sim\varepsilon\sim\delta$, at least for uniform meshes.

We perform a numerical experiment with $\omega=(-0.5,0.5)\times(-0.1,0.1)$ and $\mathbf{h}_{\text{ext}}=(1,-0.3)^T$. This choice ensures that data is smooth, no symmetries arise, and the constraint $|\mathbf{m}|\le 1$ is active on a large subdomain. Following the heuristics, we choose $\varepsilon=\delta=h$. To estimate various error quantities, we computed a reference solution on a mesh with $h=7.8125\cdot 10^{-4}$ and $1,964,544$ degrees of freedom. Figure 1 shows the results: The dominant error contribution is the error of the divergence in the V-norm $\|\nabla\cdot(\mathbf{m}^{**}-\mathbf{m}^{\delta})\|_V$ that decays at a rate of $\mathscr{O}(h^{1/2})$. The error measured in the L^2-norm is of higher order. The energy is approximated with $\mathscr{O}(h)$.

References

1. Carstensen, C., Prohl, A.: Numerical analysis of relaxed micromagnetics by penalised finite elements. Numer. Math. **90**(1), 65–99 (2001).

2. Ciarlet, P. G.: The finite element method for elliptic problems. North-Holland Publishing Co., Amsterdam, 1978.
3. DeSimone, A., Kohn, R., Müller, S., Otto, F.: A reduced theory for thin-film micromagnetics. Comm. Pure Appl. Math. **55**(11), 1408–1460 (2002)
4. DeSimone, A., Kohn, R., Müller, S., Otto, F., Schäfer, R.: Two-dimensional modelling of soft ferromagnetic films. R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci. **457**(2016), 2983–2991 (2001)
5. Ferraz-Leite, S.: Quadratic minimization with non-local operators and non-linear constraints. Dissertation, Vienna University of Technology, Vienna (2011)
6. Ferraz-Leite, S., Melenk, J.M., Praetorius, D.: Reduced Model in Thin-Film Micromagnetics. In: Troch, I., Breitenecker, F. (edts.) Proceedings MATHMOD 09 Vienna, ARGESIM Report no. 35, Vienna (2009)
7. Ferraz-Leite, S., Melenk, J.M., Praetorius, D.: Numerical quadratic energy minimization bound to convex constraints in thin-film micromagnetics. Numer. Math., In Press, Available online 2 March 2012
8. Landau, L.D., Lifschits, E.M.: On the theory of the dispersion of magnetic permeability in ferromagnetic bodies. Phys. Zeitsch. der Sow. **8**, 153–169 (1935)
9. Raviart, P.A., Thomas, J.M.: Primal hybrid finite element methods for 2nd order elliptic equations. Math. Comp. **31**(138), 391–413 (1977)
10. Sauter, A., Schwab, C.: Boundary Element Methods. Springer. Berlin (2011)
11. Stephan, E.P.: Boundary integral equations for screen problems in $\mathbf{R}^3$. Integral Equations Operator Theory **10**(2), 236–257 (1987)

Second Order Finite Volume Scheme for Maxwell's Equations with Discontinuous Dielectric Permittivity on Prismatic Meshes

T.Z. Ismagilov

Abstract A second order finite volume scheme for numerical solution of Maxwell's equations with discontinuous dielectric permittivity on prismatic meshes is suggested. The scheme is based on the approaches of Godunov, Lax-Wendoff and Van Leer. The key feature of the scheme is gradient calculation and limitation that guarantee approximation even near dielectric permittivity discontinuities. Numerical tests confirm second order of approximation of the proposed scheme.

1 Introduction

Maxwell's equations are the fundamental equations of electrodynamics. They describe propagation and diffraction of electromagnetic waves [14]. For many practical problems no analytic solution exists and, as a result, various numerical methods were developed [11].

Probably, the most popular numerical method for Maxwell's equations today is the finite difference method [15]. The first scheme of this kind was suggested in [18]. This scheme uses staggered cartesian grids and is second order accurate in space and time for the case of constant dielectric permittivity. Unfortunately cartesian grids do not allow the finite difference method to adequately represent curvilinear boundaries. The finite volume method does not have this shortcoming. It can use curvilinear structured grids [9, 12] or unstructured grids [1, 3, 8] providing better representation of curvilinear boundaries.

Arguably, the main challenge for both the finite difference and the finite volume methods is a case of discontinuous dielectric permittivity. Several approaches were suggested to address this problem. For the finite difference method various ways

T.Z. Ismagilov (✉)
Novosibirsk State University, Pirogova 2, Novosibirsk, 630090, Russia
e-mail: ismagilov@ccfit.nsu.ru

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_12,

of dielectric permittivity smoothing were discussed [16]. For the finite volume method an approach based on continuous variables was proposed [8, 13] for linear discontinuities. Later it was extended to curvilinear discontinuities [4, 5]. For some test cases the suggested approaches made it possible to achieve smaller error but were not successful in keeping the second order of approximation of the initial schemes.

In this paper we suggest a second order finite volume scheme for numerical solution of three-dimensional Maxwell's equations with discontinuous dielectric permittivity on prismatic grids. The scheme is based on the Godunov scheme [2]. To increase the order of approximation we use the approaches of Van Leer [17] and Lax-Wendroff [7] similar to [8, 9, 13]. The main difference of the suggested scheme from the schemes [4, 5, 8, 13] is approximation and limitation of spatial derivatives that is first order accurate even for the cells adjacent to the discontinuities of electric permittivity. Calculation results confirm the second order of approximation.

2 Maxwell's Equations

The system of three-dimensional Maxwell's equations can be written using non-dimensional variables in a vector form as:

$$\frac{\partial}{\partial t}\mathbf{U} + \frac{\partial}{\partial x_1}\mathbf{F}_1 + \frac{\partial}{\partial x_2}\mathbf{F}_2 + \frac{\partial}{\partial x_3}\mathbf{F}_3 = 0, \tag{1}$$

where $\mathbf{U}$ is the vector of conservative variables, $\mathbf{F}_1$, $\mathbf{F}_2$ and $\mathbf{F}_3$ are the flux vectors

$$\mathbf{U} = \begin{pmatrix} D_1 \\ D_2 \\ D_3 \\ B_1 \\ B_2 \\ B_3 \end{pmatrix}, \ \mathbf{F}_1 = \begin{pmatrix} 0 \\ H_3 \\ -H_2 \\ 0 \\ -E_3 \\ E_2 \end{pmatrix}, \ \mathbf{F}_2 = \begin{pmatrix} -H_3 \\ 0 \\ H_1 \\ E_3 \\ 0 \\ -E_1 \end{pmatrix}, \ \mathbf{F}_3 = \begin{pmatrix} H_2 \\ -H_1 \\ 0 \\ -E_2 \\ E_1 \\ 0 \end{pmatrix}.$$

In the above formulas $\mathbf{E}$ is the electric field, $\mathbf{H}$ the magnetic field, $\mathbf{D} = \varepsilon\mathbf{E}$ the electric induction, $\mathbf{B} = \mu\mathbf{H}$ the magnetic induction, ε the dielectric permittivity, and μ the magnetic permeability (assumed 1.0 in this paper). The system of Maxwell's equations can also be written using the flux variables $\mathbf{V}$ related to the conservative variables by $\mathbf{U} = Q\mathbf{V}$ as

$$Q\frac{\partial}{\partial t}\mathbf{V} + A_1\frac{\partial}{\partial x_1}\mathbf{V} + A_2\frac{\partial}{\partial x_2}\mathbf{V} + A_3\frac{\partial}{\partial x_3}\mathbf{V} = 0, \tag{2}$$

where

$$Q = \begin{pmatrix} \varepsilon & 0 & 0 & 0 & 0 & 0 \\ 0 & \varepsilon & 0 & 0 & 0 & 0 \\ 0 & 0 & \varepsilon & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{V} = \begin{pmatrix} E_1 \\ E_2 \\ E_3 \\ H_1 \\ H_2 \\ H_3 \end{pmatrix},$$

and the matrices A_1, A_2, A_3 are written as

$$A_1 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad A_3 = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

3 Numerical Scheme

By integrating the system of Maxwell's equations over a prismatic cell Δ_i with faces Γ_k assuming constant dielectric permittivity in the cell we can obtain an integral conservation law

$$Q \frac{\partial}{\partial t} \int\limits_{\Delta_i} \mathbf{V} d\Omega + \sum_{k=1}^{5} \int\limits_{\Gamma_k} (n_1 \mathbf{F}_1 + n_2 \mathbf{F}_2 + n_3 \mathbf{F}_3)\, d\Gamma = 0, \tag{3}$$

where (n_1, n_2, n_3) is a unit normal directed outside. For approximation of this equation consider a finite volume Godunov scheme

$$Q \Omega_{\Delta_i} \frac{\mathbf{V}_i^{n+1} - \mathbf{V}_i^n}{\tau} + \sum_{k=1}^{5} s_{\Delta_i}^k \mathbf{F}_i^k = 0, \tag{4}$$

where Ω_{Δ_i} is the volume of the ith cell, $s_{\Delta_i}^k$ the area of its kth face and the flux $\mathbf{F}$ is calculated using the exact solution [12] to the Riemann problem $\mathbf{F} = A^+ \mathbf{V}_L \left(\mathbf{X}^\Gamma\right) + A^- \mathbf{V}_R \left(\mathbf{X}^\Gamma\right)$. The matrices A^+ and A^- can be written using the notation $Y = \sqrt{\varepsilon}$ as

$$A^+ = \frac{1}{Y_L + Y_R}\begin{pmatrix} Y_R Y_L(1-n_1^2) & -Y_R Y_L n_1 n_2 & -Y_R Y_L n_1 n_3 & 0 & Y_R n_3 & -Y_R n_2 \\ -Y_R Y_L n_1 n_2 & -Y_R Y_L(1-n_2^2) & -Y_R Y_L n_2 n_3 & -Y_R n_3 & 0 & Y_R n_1 \\ -Y_R Y_L n_1 n_3 & -Y_R Y_L n_2 n_3 & Y_R Y_L(1-n_3^2) & Y_R n_2 & -Y_R n_1 & 0 \\ 0 & -Y_L n_3 & Y_L n_2 & 1-n_1^2 & -n_1 n_2 & -n_1 n_3 \\ Y_L n_3 & 0 & -Y_L n_1 & -n_1 n_2 & 1-n_2^2 & -n_2 n_3 \\ -Y_L n_2 & Y_L n_1 & 0 & -n_1 n_3 & -n_2 n_3 & 1-n_3^2 \end{pmatrix},$$

$$A^- = \frac{1}{Y_L + Y_R}\begin{pmatrix} Y_R Y_L(n_1^2-1) & Y_R Y_L n_1 n_2 & Y_R Y_L n_1 n_3 & 0 & Y_L n_3 & -Y_L n_2 \\ Y_R Y_L n_1 n_2 & Y_R Y_L(n_2^2-1) & Y_R Y_L n_2 n_3 & -Y_L n_3 & 0 & Y_L n_1 \\ Y_R Y_L n_1 n_3 & Y_R Y_L n_2 n_3 & Y_R Y_L(n_3^2-1) & Y_L n_2 & -Y_L n_1 & 0 \\ 0 & -Y_R n_3 & Y_R n_2 & n_1^2-1 & n_1 n_2 & n_1 n_3 \\ Y_R n_3 & 0 & -Y_R n_1 & n_1 n_2 & n_2^2-1 & n_2 n_3 \\ -Y_R n_2 & Y_R n_1 & 0 & n_1 n_3 & n_2 n_3 & n_3^2-1 \end{pmatrix}.$$

This scheme will have the second order of approximation in space and time if the values at the edge centers are approximated with the second order. Such values can be obtained using an interpolation

$$\mathbf{V}_{L,R}(\mathbf{X}^\Gamma) = \mathbf{V}(\mathbf{X}_{L,R}) + \frac{\partial \mathbf{V}}{\partial \mathbf{x}}(\mathbf{X}_{L,R})(\mathbf{X}^\Gamma - \mathbf{X}_{L,R}) - \frac{\tau}{2} Q^{-1} \sum_{j=1}^{3} A_j \frac{\partial \mathbf{V}}{\partial x_j}(\mathbf{X}_{L,R}), \quad (5)$$

if the derivatives are approximated with the first order of accuracy [7, 17].

To approximate the derivatives we will use a two-stage procedure similar to the one considered in [8, 13]. But unlike the procedure used in [8, 13] the procedure suggested here will provide approximation of the derivatives with the first order.

During the first stage in each vertex for each adjacent region with constant dielectric permittivity we will calculate the limit of the electromagnetic field with the second order of approximation. To accomplish this, in each cell we will calculate the derivatives using the values in it and the values in its neighbors with the same dielectric permittivity as in the cell with the help of the least squares method [6]. Then we will extrapolate the values in the cells on the adjacent vertices using the calculated derivatives. The limit value at a vertex will be taken as an arithmetic average of all the interpolations from the cells with the same dielectric permittivity.

During the second stage we will calculate the derivatives in each cell using the limit values at the vertices of the cell corresponding to the same value of dielectric permittivity as in the cell.

4 Computational Results

To verify the approximation properties of the proposed scheme we performed a range of test computations for a number of problems. Here we present the results for only one such problem. The results for the other problems led to the same conclusions. The prismatic meshes were built with the help of the two-dimensional

triangular meshes. The precision was evaluated by comparing the numerical solution to the analytic one. The error at the time moment $t^n = n\tau$ was evaluated using

$$\frac{\left\|\mathbf{V}^n\left(\mathbf{X}^B\right) - \mathbf{V}^{\mathbf{exact}}\left(\mathbf{X}^B, t^n\right)\right\|_{L_2}}{\left\|\mathbf{V}^{\mathbf{exact}}\left(\mathbf{X}^B, t^n\right)\right\|_{L_2}} = \sqrt{\frac{\sum\limits_{i=1}^{P}\left[\sum\limits_{k=1}^{6}\left(\mathbf{V}_k^n\left(\mathbf{X}^{B_i}\right) - \mathbf{V}_k^{\mathbf{exact}}\left(\mathbf{X}^{B_i}, t^n\right)\right)^2\right]\cdot S_{\Delta_i}}{\sum\limits_{i=1}^{P}\left[\sum\limits_{k=1}^{6}\left(\mathbf{V}_k^{\mathbf{exact}}\left(\mathbf{X}^{B_i}, t^n\right)\right)^2\right]\cdot S_{\Delta_i}}}$$

where P is the total number of prisms in the computational region, $\mathbf{V}_k^n\left(\mathbf{X}^{B_i}\right)$ and $\mathbf{V}_k^{\mathbf{exact}}\left(\mathbf{X}^{B_i}, t^n\right)$ are the computed and the exact values of electromagnetic fields at the center of the cell i.

Consider propagation of a hybrid electromagnetic mode in a waveguide with a step-like profile of the dielectric permittivity. The discontinuity of the dielectric permittivity ε is located along the curvilinear surface $r \equiv \sqrt{x_1^2 + x_2^2} = a$

$$\varepsilon = \varepsilon(r) = \begin{cases} \varepsilon_1 = n_1^2, & 0 \le r \le a, \\ \varepsilon_2 = n_2^2, & r > a. \end{cases}$$

In this case the system of Maxwell's equations has an analytic solution [10] that in the cylindric coordinates can be written in the form

$$\begin{aligned}
0 \le r \le a: E_r &= \beta\tfrac{a}{u}\left[\tfrac{1-s}{2}J_0\left(\tfrac{u}{a}r\right) - \tfrac{1+s}{2}J_2\left(\tfrac{u}{a}r\right)\right]\cos(\theta)\sin\left(kt-\beta z\right), \\
E_\theta &= -\beta\tfrac{a}{u}\left[\tfrac{1-s}{2}J_0\left(\tfrac{u}{a}r\right) + \tfrac{1+s}{2}J_2\left(\tfrac{u}{a}r\right)\right]\sin(\theta)\sin\left(kt-\beta z\right), \\
E_z &= J_1\left(\tfrac{u}{a}r\right)\cdot\cos(\theta)\cos\left(kt-\beta z\right), \\
H_r &= kn_1^2\tfrac{a}{u}\left[\tfrac{1-s_1}{2}J_0\left(\tfrac{u}{a}r\right) + \tfrac{1+s_1}{2}J_2\left(\tfrac{u}{a}r\right)\right]\sin(\theta)\sin\left(kt-\beta z\right) \\
H_\theta &= kn_1^2\tfrac{a}{u}\left[\tfrac{1-s_1}{2}J_0\left(\tfrac{u}{a}r\right) - \tfrac{1+s_1}{2}J_2\left(\tfrac{u}{a}r\right)\right]\cos(\theta)\sin\left(kt-\beta z\right), \\
H_z &= -\tfrac{\beta}{k}sJ_1\left(\tfrac{u}{a}r\right)\cdot\sin(\theta)\cos\left(kt-\beta z\right),
\end{aligned}$$

$$\begin{aligned}
r > a: E_r &= \beta\tfrac{a}{w}\tfrac{J_1(u)}{K_1(w)}\left[\tfrac{1-s}{2}K_0\left(\tfrac{w}{a}r\right) - \tfrac{1+s}{2}K_2\left(\tfrac{w}{a}r\right)\right]\cos(\theta)\sin\left(kt-\beta z\right), \\
E_\theta &= -\beta\tfrac{a}{w}\tfrac{J_1(u)}{K_1(w)}\left[\tfrac{1-s}{2}K_0\left(\tfrac{w}{a}r\right) + \tfrac{1+s}{2}K_2\left(\tfrac{w}{a}r\right)\right]\sin(\theta)\sin\left(kt-\beta z\right), \\
E_z &= \tfrac{J_1(u)}{K_1(w)}K_1\left(\tfrac{w}{a}r\right)\cdot\cos(\theta)\cos\left(kt-\beta z\right),
\end{aligned}$$

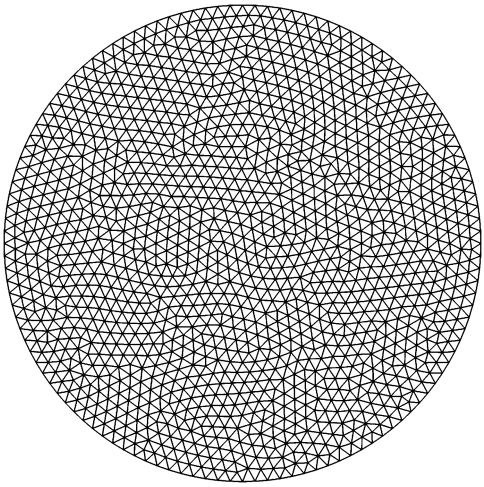


Fig. 1 Two-dimensional mesh used to generate the three-dimensional prismatic mesh

$$H_r = k n_0^2 \frac{a}{w} \frac{J_1(u)}{K_1(w)} \left[\frac{1-s_0}{2} K_0\left(\frac{w}{a} r\right) + \frac{1+s_0}{2} K_2\left(\frac{w}{a} r\right) \right] \sin(\theta) \sin\left(k t - \beta z\right)$$

$$H_\theta = k n_0^2 \frac{a}{w} \frac{J_1(u)}{K_1(w)} \left[\frac{1-s_0}{2} K_0\left(\frac{u}{a} r\right) - \frac{1+s_0}{2} K_2\left(\frac{w}{a} r\right) \right] \cos(\theta) \sin\left(k t - \beta z\right),$$

$$H_z = -\frac{\beta}{k} s \frac{J_1(u)}{K_1(w)} K_1\left(\frac{w}{a} r\right) \cdot \sin(\theta) \cos\left(k t - \beta z\right),$$

where J_0 and J_1 are Bessel functions of the first kind, K_0 and K_1 Bessel functions of the second kind, $s_0 = s\beta^2/k^2 n_0^2$, $s_1 = s\beta^2/k^2 n_1^2$, $u = a\sqrt{k^2 n_1^2 - \beta^2}$, $w = a\sqrt{k^2 n_2^2 - \beta^2}$,

$$s = 2\left(\frac{1}{u^2} + \frac{1}{w^2}\right) \left[\frac{J_0(u) - J_2(u)}{u J_1(u)} - \frac{K_0(w) - K_2(w)}{w K_1(w)} \right]^{-1},$$

and β can be obtained from the dispersion relation. We used the following test constants $\varepsilon_1 = 2.25$, $\varepsilon_2 = 1.0$, $k = 6.0$, $a = 0.64$, $\beta = 8.402440923258$, $u = 2.063837416842$, $w = 3.764648073438$.

As a computational region we used a cylinder of radius 1.28 and height 0.6. The computations were performed using a sequence of grids composed of 27,080, 75,964, 217,000 and 609,000 prisms. The mesh was built in such a way so that the discontinuity of dielectric permittivity was along the prism faces. The time steps were chosen proportional to the linear prism sizes. Figure 1 shows a two-dimensional triangular mesh composed of 2,708 triangles that was used to build a

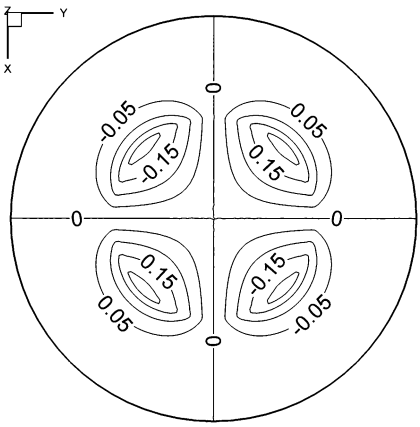


Fig. 2 Distribution of H_1 at time = 2.2 and z = 0.3

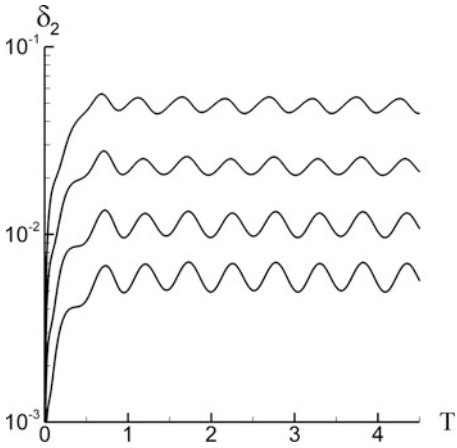


Fig. 3 Error evolution

Table 1 Maximum error

Cells	δ_2	Order
27,080	0.0560509	
75,964	0.0278387	2.01
217,000	0.0134622	2.05
609,000	0.0071236	2.00

mesh of 27,080 prisms. Figure 2 shows the distribution of the first component of the magnetic field H_1 at the time $T = 2.2$ for the cross-section $x_3 = 0.3$ obtained using a mesh of 609,000 prisms. The error evolution for a sequence of four meshes is presented in Fig. 3. Table 1 shows the maximum errors in the L_2 norm. The error behavior corresponds to the second order of approximation.

5 Conclusion

A finite volume scheme for numerical solution of Maxwell's equations with dielectric permittivity on prismatic meshes was suggested. The scheme is second order accurate in space and time. A range of numerical tests for linear as well as curvilinear discontinuities confirm the second order of approximation.

Acknowledgements The author would like to thank A.S. Lebedev for discussions on many aspects of this work.

References

1. Cioni, J.-P., Fezoui, L., Steve, H.: A Parallel Time-Domain Maxwell Solver Using Upwind Schemes and Triangular Meshes. IMPACT Comput. Sci. Eng. **5**, 215–247 (1993)
2. Godunov, S.K.: A Difference Scheme for Numerical Solution of Discontinuos Solution of Hydrodynamic Equations. Math. Sbornik **47**, 271–306 (1959)
3. Hermeline, F.: Two Coupled Particle-Finite Volume Methods Using Dalaunay-Voronoi Meshes for Approximation of Vlasov-Poisson and Vlasov-Maxwell Equations. J. Comput. Phys. **106**, 1–18 (1993)
4. Ismagilov, T.Z.: Parallel Algorithm for Numerical Solution of Three-dimensional Maxwell's Equations with Discontinuous Dielectric Permittivity on Tetrahedral Meshes (In Russian). USATU Vestnik **14**, 152–159 (2010)
5. Ismagilov, T.Z., Gorbachev, A.I.: Parallel Algorithm for Numerical Solution of Three-dimensional Maxwell's Equations with Discontinuous Dielectric Permittivity on Prismatic Meshes (In Russian). Comp. Meth. and Progr. **12**, 128–136 (2011)
6. Laub, A.J.: Matrix Analysis for Scientists and Engineers. SIAM, Philadelphia (2004)
7. Lax, P.D., Wendroff, B.: Systems of Conservation Laws. Commun. Pure Appl. Math. **13**, 217–237 (1960)
8. Lebedev, A.S., Fedoruk, M.P., Shtyrina, O.V.: Finite-Volume Algorithm for Solving the Time-Dependent Maxwell Equations on Unstructured Meshes. Comp. Math. and Math. Phys. **47**, 1286–1301 (2006)
9. Munz, C.-D., Schneider, R., Voss, U.: A Finite-Volume Method for the Maxwell Equations in the Time Domain. SIAM J. on Sci. Comp. **22**, 449–475 (2000)
10. Okamoto, K.: Fundamentals of Optical Waveguides. Academic Press, London (2000)
11. Rao, S.M.: Time Domain Electromagnetics. Academic Press, San Diego (1999)
12. Shankar, V., Hall, W.F., Mohammadian, A.H.: A CFD-based Finite-Volume Procedure for Computational Electromagnetics - Interdisciplinary Applications of CFD Methods. AIAA A89-41776 **18–02**, 551–564 (1989)
13. Shokin, Yu.I., Lebedev, A.S., Shtyrina, O.V., Fedoruk, M.P.: Solution of Maxwell's equations on partially unstructured meshes. Notes on Numerical Fluid Mechanics and Multidisciplinary Design **93**, 1–13 (2005)
14. Stratton, J.A.: Electromagnetic Theory. McGraw Hill, NewYork (1941)
15. Sullivan, D.M.: Electromagnetic Simulation Using the Finite-Difference Time-Domain Method. Wiley-IEEE Press (2000)
16. Taflove, A., Hagness, S.C.: Computational Electrodynamics: the Finite-Difference Time-Domain Method. Artech House, Norwood (2005)
17. Van Leer, B.: Towards the Ultimate Conservative Difference Scheme. A Second Order Sequel to Godunov's Method. J. Comput. Phys. **32**, 101–136 (1979)
18. Yee, K.S.: Numerical Solution of Initial Boundary Value Problems Involving Maxwell's Equations in Isotropic Media. IEEE Trans. Antennas Propagat. **14**, 585–589 (1966)

A Hybridizable Discontinuous Galerkin Method for Solving 3D Time-Harmonic Maxwell's Equations

L. Li, S. Lanteri, and R. Perrussel

Abstract We study the numerical solution of 3d time-harmonic Maxwell's equations by a hybridizable discontinuous Galerkin method. A hybrid term representing the tangential component of the numerical trace of the magnetic field is introduced. The global system to solve only involves the hybrid term as unknown. We show that the reduced system has properties similar to wave equation discretizations and the tangential components of the numerical traces for both electric and magnetic fields are single-valued. On the example of a plane wave propagation in vacuum the approximate solutions for both electric and magnetic fields have an optimal convergence order.

1 Introduction

Discontinuous Galerkin (DG) methods have recently been considered for the approximate solution of Maxwell's equations, see [2, 4, 5, 7]. Thanks to the discontinuity, this kind of methods has many advantages, such as adaptivity to complex geometries, easily obtained high order accuracy, hp-adaptivity and natural

L. Li (✉)
School of Mathematical Sciences, University of Electronic Science and Technology of China, Chengdu, Sichuan, 610054, P.R. China (Liang Li is also with INRIA, NACHOS project-team)
e-mail: liang.li@inria.fr; plum.liliang@gmail.com

S. Lanteri
INRIA, NACHOS project-team, 2004 Route des Lucioles, B.P. 93, 06902 Sophia Antipolis Cedex, France
e-mail: Stephane.Lanteri@inria.fr

R. Perrussel
Université de Toulouse, CNRS/INPT/UPS, LAPLACE (LAboratoire PLasma et Conversion d'Énergie), Toulouse, France
e-mail: perrussel@laplace.univ-tlse.fr

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_13,

parallelism [6]. Despite many advantages, the DG methods have one main drawback particularly sensitive for stationary problems: the number of globally coupled degrees of freedom is "much" greater than the number of degrees of freedom used by conforming finite element methods for the same accuracy. Consequently, the DG methods are expensive in terms of both CPU time and memory consumption, especially for time-harmonic problems [3]. Hybridization of DG methods [1] is devoted to address this issue while keeping all the advantages of DG methods.

The hybridizable discontinuous Galerkin (HDG) methods introduce an additional *hybrid* variable on the faces of the element, with which the local solutions can be defined. A so-called *conservativity condition* is imposed on the numerical traces, which can be represented by the hybrid variable at the interface between neighboring elements [1, 8]. As a result, the HDG methods produce a linear system in terms of the degrees of freedom of the additional hybrid variable only. In this way, the number of globally coupled degrees of freedom is reduced. The local solutions of the electromagnetic fields can be recovered by solving local problems element-by-element. For 3D Maxwell's equations, we propose a HDG formulation taking the tangential component of the magnetic field as the hybrid variable. We show that the reduced system of the hybrid variable has a wave-equation-like characterization and the tangential components of the numerical traces for both electric and magnetic fields are single-valued. Moreover, in our numerical examples the approximate solutions for both electric and magnetic fields have an optimal convergence order.

The outline of this paper is as follows. In Sect. 2, we present 3D Maxwell's equations and give some notations. In Sect. 3, we propose a HDG formulation and discuss the characterization of the global reduced problem and the continuity of the numerical traces on the interfaces. Section 4 contains our preliminary numerical results. We give a brief conclusion in Sect. 5.

2 Problem Statement and Notations

2.1 Time-Harmonic Maxwell's Equations in 3D

Time-harmonic Maxwell's equations without volume source are considered

$$\begin{cases} \mathrm{i}\omega\varepsilon_r \mathbf{E} - \mathbf{curl}\,\mathbf{H} = 0, \text{ in } \Omega, \\ \mathrm{i}\omega\mu_r \mathbf{H} + \mathbf{curl}\,\mathbf{E} = 0, \text{ in } \Omega, \\ \mathbf{n}\times\mathbf{E} = 0, \text{ on } \Gamma_m, \\ \mathbf{n}\times\mathbf{E} + \mathbf{n}\times(\mathbf{n}\times\mathbf{H}) = \mathbf{n}\times\mathbf{E}^{\mathrm{inc}} + \mathbf{n}\times(\mathbf{n}\times\mathbf{H}^{\mathrm{inc}}), \text{ on } \Gamma_a, \end{cases} \tag{1}$$

where i is the imaginary unit, ω is the angular frequency, ε_r and μ_r are the relative permittivity and permeability respectively, $\mathbf{n}$ is the outgoing normal, $\mathbf{E}$ and $\mathbf{H}$ are the

electric and magnetic fields, and $(\mathbf{E}^{\text{inc}}, \mathbf{H}^{\text{inc}})$ is the incident electromagnetic wave. The boundary of the computational domain Ω is $\partial\Omega = \Gamma_m \cup \Gamma_a$.

2.2 Notations

We consider a simplicial mesh $\mathscr{T}_h$ (consists of tetrahedrons) of the computational domain Ω. We denote by $\mathscr{F}_h^I$ the union of all interior interfaces of $\mathscr{T}_h$, by $\mathscr{F}_h^B$ the union of all the boundary interfaces of $\mathscr{T}_h$, and by $\mathscr{F}_h = \mathscr{F}_h^I \cup \mathscr{F}_h^B$. For an interface $F = \overline{K^+} \cap \overline{K^-} \in \mathscr{F}_h^I$, let $\mathbf{v}^\pm$ be the traces of $\mathbf{v}$ on F from the interior of $K^\pm$. On this face, we define *mean* (*average*) *values* $\{\cdot\}$ and *jumps* $[\![\cdot]\!]$ as follows:

$$\begin{cases} \{\mathbf{v}\}_F = \dfrac{1}{2}(\mathbf{v}^+ + \mathbf{v}^-), \\ [\![\mathbf{v}]\!]_F = \mathbf{n}_{K^+} \times \mathbf{v}^+ + \mathbf{n}_{K^-} \times \mathbf{v}^-, \end{cases}$$

where $\mathbf{n}_{K^\pm}$ is the outgoing normal of $K^\pm$. For the boundary faces these expressions are turned to be

$$\begin{cases} \{\mathbf{v}\}_F = \mathbf{v}^+, \\ [\![\mathbf{v}]\!]_F = \mathbf{n}_{K^+} \times \mathbf{v}^+, \end{cases}$$

since we assume $\mathbf{v}$ is single-valued on the boundaries. Let $\mathbb{P}_p(D)$ denote the space of polynomial functions of degree at most p on a domain D. For any element $K \in \mathscr{T}_h$, let $\mathbf{V}^p(K) := (\mathbb{P}_p(K))^3$. The discontinuous finite element spaces are then introduced by

$$\mathbf{V}_h^p = \left\{ \mathbf{v} \in (L^2(\Omega))^3 \mid \mathbf{v}|_K \in \mathbf{V}^p(K),\ \forall K \in \mathscr{T}_h \right\},$$

where $L^2(\Omega)$ is the space of square-integrable functions on the domain Ω. We introduce a traced finite element space

$$\mathbf{M}_h^p = \left\{ \boldsymbol{\eta} \in (L^2(\mathscr{F}_h))^3 \mid \boldsymbol{\eta}|_F \in (\mathbb{P}_p(F))^3, (\boldsymbol{\eta} \cdot \mathbf{n})|_F = 0\ \forall F \in \mathscr{F}_h \right\}. \tag{2}$$

Note that $\mathbf{M}_h^p$ consists of vector-valued functions whose normal component is zero on any face $F \in \mathscr{F}_h$. For two vector functions $\mathbf{u}$ and $\mathbf{v}$ in $(L^2(D))^3$, we denote $(\mathbf{u}, \mathbf{v})_D = \int_D \mathbf{u} \cdot \mathbf{v}\, \mathrm{d}x$ provided D is a domain in $\mathbb{R}^3$, and we denote $\langle \mathbf{u}, \mathbf{v} \rangle_F = \int_F \mathbf{u} \cdot \mathbf{v}\, \mathrm{d}s$ if F is a two-dimensional face. Accordingly, for the mesh $\mathscr{T}_h$ we have

$$(\cdot,\cdot)_{\mathscr{T}_h} = \sum_{K \in \mathscr{T}_h} (\cdot,\cdot)_K, \qquad \langle \cdot,\cdot \rangle_{\partial\mathscr{T}_h} = \sum_{K \in \mathscr{T}_h} \langle \cdot,\cdot \rangle_{\partial K},$$

$$\langle \cdot,\cdot \rangle_{\mathscr{F}_h} = \sum_{F \in \mathscr{F}_h} \langle \cdot,\cdot \rangle_F, \qquad \langle \cdot,\cdot \rangle_{\Gamma_a} = \sum_{F \in \mathscr{F}_h \cap \Gamma_a} \langle \cdot,\cdot \rangle_F.$$

We set

$$\mathbf{v}^t = -\mathbf{n} \times (\mathbf{n} \times \mathbf{v}), \quad \mathbf{v}^n = \mathbf{n}(\mathbf{n} \cdot \mathbf{v}),$$

where $\mathbf{v}^t$ and $\mathbf{v}^n$ denote the tangential and normal components of $\mathbf{v}$, and $\mathbf{v} = \mathbf{v}^t + \mathbf{v}^n$.

3 HDG Formulation

3.1 Principles

The discontinuous Galerkin (DG) method seeks an approximate solution $(\mathbf{E}_h, \mathbf{H}_h)$ in the space $\mathbf{V}_h^p \times \mathbf{V}_h^p$ whose objective is to satisfy for all K in $\mathscr{T}_h$

$$\begin{cases} (\mathrm{i}\omega\varepsilon_r \mathbf{E}_h, \overline{\mathbf{v}})_K - (\mathbf{curl}\, \mathbf{H}_h, \overline{\mathbf{v}})_K = 0, \ \forall \mathbf{v} \in \mathbf{V}^p(K), \\ (\mathrm{i}\omega\mu_r \mathbf{H}_h, \overline{\mathbf{v}})_K + (\mathbf{curl}\, \mathbf{E}_h, \overline{\mathbf{v}})_K = 0, \ \forall \mathbf{v} \in \mathbf{V}^p(K). \end{cases} \tag{3}$$

The *numerical traces* $\widehat{\mathbf{E}}_h$ and $\widehat{\mathbf{H}}_h$ are introduced through integration by parts

$$\begin{cases} (\mathrm{i}\omega\varepsilon_r \mathbf{E}_h, \overline{\mathbf{v}})_K - (\mathbf{H}_h, \overline{\mathbf{curl}\, \mathbf{v}})_K + \langle \widehat{\mathbf{H}}_h^t, \overline{\mathbf{n} \times \mathbf{v}} \rangle_{\partial K} = 0, \ \forall \mathbf{v} \in \mathbf{V}^p(K), \\ (\mathrm{i}\omega\mu_r \mathbf{H}_h, \overline{\mathbf{v}})_K + (\mathbf{E}_h, \overline{\mathbf{curl}\, \mathbf{v}})_K - \langle \widehat{\mathbf{E}}_h, \overline{\mathbf{n} \times \mathbf{v}} \rangle_{\partial K} = 0, \ \forall \mathbf{v} \in \mathbf{V}^p(K). \end{cases} \tag{4}$$

In a classic DG method, we directly define the numerical traces to couple the neighboring traces. In HDG method, we introduce a hybrid term $\boldsymbol{\Lambda}_h$,

$$\boldsymbol{\Lambda}_h := \widehat{\mathbf{H}}_h^t, \quad \forall F \in \mathscr{F}_h, \tag{5}$$

and we want to solve the local fields through (4) as long as $\boldsymbol{\Lambda}_h$ is known on all the faces of an element K. In order to achieve this, we consider the numerical trace $\widehat{\mathbf{E}}_h$ of the form

$$\widehat{\mathbf{E}}_h = \mathbf{E}_h + \tau_K \mathbf{n} \times (\boldsymbol{\Lambda}_h - \mathbf{H}_h^t) \text{ on } \partial K, \tag{6}$$

where τ_K is a local stabilization parameter. Note that $\mathbf{n} \times \mathbf{H}_h^t = \mathbf{n} \times \mathbf{H}_h$. Adding the contributions of (4) over all elements and enforcing the continuity of the tangential component of $\widehat{\mathbf{E}}_h$, we can formulate a problem which is to find $(\mathbf{E}_h, \mathbf{H}_h, \boldsymbol{\Lambda}_h) \in \mathbf{V}_h^p \times \mathbf{V}_h^p \times \mathbf{M}_h^p$ such that

$$\begin{cases} (\mathrm{i}\omega\varepsilon_r \mathbf{E}_h, \overline{\mathbf{v}})_{\mathscr{T}_h} - (\mathbf{H}_h, \overline{\mathbf{curl}\, \mathbf{v}})_{\mathscr{T}_h} + \langle \boldsymbol{\Lambda}_h, \overline{\mathbf{n} \times \mathbf{v}} \rangle_{\partial \mathscr{T}_h} = 0, \ \forall \mathbf{v} \in \mathbf{V}_h^p, \\ (\mathrm{i}\omega\mu_r \mathbf{H}_h, \overline{\mathbf{v}})_{\mathscr{T}_h} + (\mathbf{E}_h, \overline{\mathbf{curl}\, \mathbf{v}})_{\mathscr{T}_h} - \langle \widehat{\mathbf{E}}_h, \overline{\mathbf{n} \times \mathbf{v}} \rangle_{\partial \mathscr{T}_h} = 0, \ \forall \mathbf{v} \in \mathbf{V}_h^p, \\ \langle [\![\mathbf{n} \times \widehat{\mathbf{E}}_h]\!], \overline{\boldsymbol{\eta}} \rangle_{\mathscr{F}_h} - \langle \boldsymbol{\Lambda}_h, \overline{\boldsymbol{\eta}} \rangle_{\Gamma_a} = \langle \mathbf{g}^{\mathrm{inc}}, \overline{\boldsymbol{\eta}} \rangle_{\Gamma_a}, \ \forall \boldsymbol{\eta} \in \mathbf{M}_h^p, \end{cases} \tag{7}$$

where the last equation is called the *conservativity condition*.

With the definition (6) of $\widehat{\mathbf{E}}_h$, we employ integration by parts again in the second equation of (7), then we get the following formulation

$$\begin{cases} (\mathrm{i}\omega\varepsilon_r\mathbf{E}_h,\overline{\mathbf{v}})_{\mathscr{T}_h} - (\mathbf{H}_h,\overline{\mathbf{curl}\,\mathbf{v}})_{\mathscr{T}_h} + \langle \boldsymbol{\Lambda}_h,\overline{\mathbf{n}\times\mathbf{v}}\rangle_{\partial\mathscr{T}_h} = 0,\ \forall\mathbf{v}\in\mathbf{V}_h^p, \\ (\mathrm{i}\omega\mu_r\mathbf{H}_h,\overline{\mathbf{v}})_{\mathscr{T}_h} + (\mathbf{curl}\,\mathbf{E}_h,\overline{\mathbf{v}})_{\mathscr{T}_h} + \langle \tau\mathbf{n}\times(\mathbf{H}_h-\boldsymbol{\Lambda}_h),\overline{\mathbf{n}\times\mathbf{v}}\rangle_{\partial\mathscr{T}_h} = 0, \\ \qquad\qquad \forall\mathbf{v}\in\mathbf{V}_h^p, \\ \langle \mathbf{n}\times\mathbf{E}_h,\overline{\boldsymbol{\eta}}\rangle_{\partial\mathscr{T}_h} + \langle \tau(\mathbf{H}_h^t-\boldsymbol{\Lambda}_h),\overline{\boldsymbol{\eta}}\rangle_{\partial\mathscr{T}_h} - \langle \boldsymbol{\Lambda}_h,\overline{\boldsymbol{\eta}}\rangle_{\Gamma_a} \\ \qquad\qquad = \langle \mathbf{g}^{\mathrm{inc}},\overline{\boldsymbol{\eta}}\rangle_{\Gamma_a},\ \forall\boldsymbol{\eta}\in\mathbf{M}_h^p. \end{cases} \tag{8}$$

Note that we have used the following relations to obtain (8)

$$\mathbf{n}\times\mathbf{v}^t = \mathbf{n}\times\mathbf{v} \text{ and } \mathbf{n}\times(\mathbf{n}\times\mathbf{v}^t) = -\mathbf{v}^t.$$

Remark 1. Assuming the interpolation order is p on every element, we obtain the number of the globally coupled degrees of freedom

$$\text{DG-}\mathbb{P}_p\text{"upwind flux" formulation (see [3])}: \quad (p+1)(p+2)(p+3)N_e,$$
$$\text{HDG-}\mathbb{P}_p: \quad (p+1)(p+2)N_f,$$

with N_e be the number of elements and N_f the number of faces. For a simplicial mesh, $N_f \approx 2N_e$, so the ratio of the globally coupled degrees of freedom of HDG method over DG method is roughly $\frac{2}{p+3}$.

3.2 Characterization of the Reduced Problem

For any $\boldsymbol{\eta}\in\mathbf{M}_h$, we denote by $(\mathbf{E}_h^{\boldsymbol{\eta}},\mathbf{H}_h^{\boldsymbol{\eta}})$ the fields whose restriction to an element K of $\mathscr{T}_h$ is the solution of the local problem:

$$\begin{cases} (\mathrm{i}\omega\varepsilon_r\mathbf{E}_h^{\boldsymbol{\eta}},\overline{\mathbf{v}})_K - (\mathbf{H}_h^{\boldsymbol{\eta}},\overline{\mathbf{curl}\,\mathbf{v}})_K + \langle \boldsymbol{\eta},\overline{\mathbf{n}\times\mathbf{v}}\rangle_{\partial K} = 0,\ \forall\mathbf{v}\in\mathbf{V}_h^p(K), \\ (\mathrm{i}\omega\mu_r\mathbf{H}_h^{\boldsymbol{\eta}},\overline{\mathbf{v}})_K + (\mathbf{curl}\,\mathbf{E}_h^{\boldsymbol{\eta}},\overline{\mathbf{v}})_K + \langle \tau\mathbf{n}\times(\mathbf{H}_h^{\boldsymbol{\eta}}-\boldsymbol{\eta}),\overline{\mathbf{n}\times\mathbf{v}}\rangle_{\partial K} = 0,\ \forall\mathbf{v}\in\mathbf{V}_h^p(K). \end{cases} \tag{9}$$

We rewrite the reduced problem

$$a_h(\boldsymbol{\Lambda}_h,\boldsymbol{\eta}) = b_h(\boldsymbol{\eta}),\ \forall\boldsymbol{\eta}\in\mathbf{M}_h^p, \tag{10}$$

with

$$\begin{cases} a_h(\boldsymbol{\Lambda}_h,\boldsymbol{\eta}) = \langle [\![\mathbf{n}\times\widehat{\mathbf{E}}^{\boldsymbol{\Lambda}_h}]\!],\overline{\eta}\rangle_{\mathscr{F}_h} + \langle \boldsymbol{\Lambda}_h,\overline{\boldsymbol{\eta}}\rangle_{\Gamma_a}, \\ b_h(\boldsymbol{\eta}) = \langle \mathbf{g}^{\mathrm{inc}},\overline{\boldsymbol{\eta}}\rangle_{\Gamma_a}. \end{cases}$$

Summing the contributions of (9) over all the elements of $\mathscr{T}_h$, we obtain the following formulations by recalling the definition of $[\![\cdot]\!]$

$$\begin{cases} -(\mathrm{i}\omega\varepsilon_r\overline{\mathbf{E}_h^{\boldsymbol{\eta}}}, \mathbf{v})_{\mathscr{T}_h} - (\overline{\mathbf{H}_h^{\boldsymbol{\eta}}}, \mathbf{curl}\,\mathbf{v})_{\mathscr{T}_h} + \langle\overline{\boldsymbol{\eta}}, [\![\mathbf{n}\times\mathbf{v}]\!]\rangle_{\mathscr{F}_h} = 0, \\ (\mathrm{i}\omega\mu_r\mathbf{H}_h^{\boldsymbol{\eta}}, \overline{\mathbf{v}})_{\mathscr{T}_h} + (\mathbf{curl}\,\mathbf{E}_h^{\boldsymbol{\eta}}, \overline{\mathbf{v}})_{\mathscr{T}_h} + \langle\tau[\![\mathbf{n}\times(\mathbf{n}\times(\boldsymbol{\eta}-\mathbf{H}_h^{\boldsymbol{\eta}}))]\!], \overline{\mathbf{v}}\rangle_{\mathscr{F}_h} = 0. \end{cases} \tag{11}$$

Note that the first relation of (11) is obtained by taking the summation in the complex conjugation of the first relation of (9).

The sesquilinear form in (10) can be made explicit

$$\begin{aligned} a_h(\boldsymbol{\Lambda}_h, \boldsymbol{\eta}) &= \langle[\![\mathbf{n}\times\widehat{\mathbf{E}}_h^{\boldsymbol{\Lambda}_h}]\!], \overline{\boldsymbol{\eta}}\rangle_{\mathscr{F}_h} - \langle\boldsymbol{\Lambda}_h, \overline{\boldsymbol{\eta}}\rangle_{\Gamma_a} \\ &= \langle[\![\mathbf{n}\times\mathbf{E}_h^{\boldsymbol{\Lambda}_h}]\!], \overline{\boldsymbol{\eta}}\rangle_{\mathscr{F}_h} + \langle[\![\mathbf{n}\times(\widehat{\mathbf{E}}_h^{\boldsymbol{\Lambda}_h} - \mathbf{E}_h^{\boldsymbol{\Lambda}_h})]\!], \overline{\boldsymbol{\eta}}\rangle_{\mathscr{F}_h} - \langle\boldsymbol{\Lambda}_h, \overline{\boldsymbol{\eta}}\rangle_{\Gamma_a} \\ &= (\mathrm{i}\omega\varepsilon_r\overline{\mathbf{E}_h^{\boldsymbol{\eta}}}, \mathbf{E}_h^{\boldsymbol{\Lambda}_h})_{\mathscr{T}_h} + (\overline{\mathbf{H}_h^{\boldsymbol{\eta}}}, \mathbf{curl}\,\mathbf{E}_h^{\boldsymbol{\Lambda}_h})_{\mathscr{T}_h} + \langle[\![\mathbf{n}\times(\widehat{\mathbf{E}}_h^{\boldsymbol{\Lambda}_h} - \mathbf{E}_h^{\boldsymbol{\Lambda}_h})]\!], \overline{\boldsymbol{\eta}}\rangle_{\mathscr{F}_h} \\ &\quad - \langle\boldsymbol{\Lambda}_h, \overline{\boldsymbol{\eta}}\rangle_{\Gamma_a} (\text{by the first relation of (11), taking } \mathbf{v} = \mathbf{E}_h^{\boldsymbol{\Lambda}_h}) \\ &= (\mathrm{i}\omega\varepsilon_r\mathbf{E}_h^{\boldsymbol{\Lambda}_h}, \overline{\mathbf{E}_h^{\boldsymbol{\eta}}})_{\mathscr{T}_h} - (\mathrm{i}\omega\mu_r\mathbf{H}_h^{\boldsymbol{\Lambda}_h}, \overline{\mathbf{H}_h^{\boldsymbol{\eta}}})_{\mathscr{T}_h} - \langle\boldsymbol{\Lambda}_h, \overline{\boldsymbol{\eta}}\rangle_{\Gamma_a} \\ &\quad + \langle[\![\mathbf{n}\times(\mathbf{n}\times(\boldsymbol{\Lambda}_h - \mathbf{H}_h^{\boldsymbol{\Lambda}_h}))]\!], \overline{(\boldsymbol{\eta}-\mathbf{H}_h^{\boldsymbol{\eta}})}\rangle_{\mathscr{F}_h}, \\ &\quad (\text{by the second relation of (11), and considering the definition (6) of } \widehat{\mathbf{E}}_h) \\ &= (\mathrm{i}\omega\varepsilon_r\mathbf{E}_h^{\boldsymbol{\Lambda}_h}, \overline{\mathbf{E}_h^{\boldsymbol{\eta}}})_{\mathscr{T}_h} - (\mathrm{i}\omega\mu_r\mathbf{H}_h^{\boldsymbol{\Lambda}_h}, \overline{\mathbf{H}_h^{\boldsymbol{\eta}}})_{\mathscr{T}_h} - \langle\boldsymbol{\Lambda}_h, \overline{\boldsymbol{\eta}}\rangle_{\Gamma_a} \\ &\quad - \langle\mathbf{n}\times(\boldsymbol{\Lambda}_h - \mathbf{H}_h^{\boldsymbol{\Lambda}_h}), \overline{\mathbf{n}\times(\boldsymbol{\eta}-\mathbf{H}_h^{\boldsymbol{\eta}})}\rangle_{\partial\mathscr{T}_h}. \end{aligned}$$

The reduced operator has a wave-equation-like characterization. When ε_r and μ_r are real-valued, we can infer that the coefficient matrix $\mathbf{K}$ of the corresponding linear system is complex symmetric and all the eigenvalues lie in the left half-plane of the complex plane, since we assume real-valued basis functions. Moreover, the first two terms define the imaginary part of $\mathbf{K}$ if ε_r, μ_r and the basis functions are real-valued. This matrix is similar to the discretization of the wave equation: it is symmetric but indefinite as soon as ω is sufficiently large.

3.3 On Numerical Traces

The conservativity condition holds on all the interior faces, which means $\langle[\![\mathbf{n}\times\widehat{\mathbf{E}}_h]\!], \overline{\eta}\rangle_{\mathscr{F}_h^I} = 0$. From the choice of spaces with p constant, we can infer that $[\![\mathbf{n}\times\widehat{\mathbf{E}}_h]\!] = 0$ on every interior face on a conforming mesh. Substituting $\widehat{\mathbf{E}}_h$ with the expression in (6), we have

$$[\![\mathbf{n}\times(\mathbf{E}_h+\tau\mathbf{n}\times(\boldsymbol{\Lambda}_h-\mathbf{H}_h^t))]\!] = [\![\mathbf{n}\times\mathbf{E}_h]\!]+\tau_{K+}\mathbf{H}_h^{t+}+\tau_{K-}\mathbf{H}_h^{t-}-(\tau_{K+}+\tau_{K-})\boldsymbol{\Lambda}_h$$
$$= 0 \text{ on } \mathscr{F}_h^I.$$

Solving for $\boldsymbol{\Lambda}_h$, we obtain (assuming $\tau_{K+}+\tau_{K-}\neq 0$)

$$\widehat{\mathbf{H}}_h^t = \boldsymbol{\Lambda}_h = \frac{1}{\tau_{K+}+\tau_{K-}}(\tau_{K+}\mathbf{H}_h^{t+}+\tau_{K-}\mathbf{H}_h^{t-})+\frac{1}{\tau_{K+}+\tau_{K-}}[\![\mathbf{n}\times\mathbf{E}_h]\!] \text{ on } F. \quad (12)$$

Substituting (12) into (6), and taking the tangential components on both sides, we have

$$\widehat{\mathbf{E}}_h^t = \widehat{\mathbf{E}}_h^{t+} = \widehat{\mathbf{E}}_h^{t-} = \frac{1}{\tau_{K+}+\tau_{K-}}(\tau_{K-}\mathbf{E}_h^{t+}+\tau_{K+}\mathbf{E}_h^{t-})+\frac{\tau_{K+}\tau_{K-}}{\tau_{K+}+\tau_{K-}}[\![\mathbf{n}\times\mathbf{H}_h]\!] \text{ on } F. \quad (13)$$

Thus, the tangential components of the numerical traces for both $\widehat{\mathbf{E}}$ and $\widehat{\mathbf{H}}$ fields are single-valued.

4 Numerical Results

We consider the propagation of a plane wave in vacuum. The computational domain is chosen to be the unit cube $\Omega =]-0.5;0.5[^3$, and the Silver Müller absorbing boundary condition is imposed on the whole boundary. The electromagnetic parameters ε_r and μ_r are set to be 1 everywhere, and the angular frequency $\omega = 2\pi$. Parameter τ is set to be 1. All the results are obtained by Matlab codes.

A series of regular tetrahedral meshes are employed, which divide the unit cube into many little cubes and then each little cube is divided into six tetrahedrons. Table 1 gives the errors and convergence behaviors of both HDG-$\mathbb{P}_1$ and HDG-$\mathbb{P}_2$ methods. In Table 1, mesh size denotes the edge length of the tetrahedrons on the edges of the unit cube. We observe that the asymptotic convergence rates of the approximate solutions in L^2-norm for both $\mathbf{E}$ and $\mathbf{H}$ are optimal.

In Fig. 1, we show the eigenvalue distribution of the reduced system for both HDG-$\mathbb{P}_1$ and HDG-$\mathbb{P}_2$ methods on the coarsest mesh. We can see that all the eigenvalues are located on the left side of the imaginary axis. It can be noticed that the number of eigenvalues with a positive imaginary part increases when ω increases for a fixed discretization; this is due to the indefinite and wave-equation-like nature of the imaginary part of the matrix as underlined in Sect. 3.2.

Table 1 Convergence results of HDG methods

		HDG-$\mathbb{P}_1$					HDG-$\mathbb{P}_2$			
		$\|\mathrm{E}-\mathrm{E}_h\|_{L^2}$		$\|\mathrm{H}-\mathrm{H}_h\|_{L^2}$			$\|\mathrm{E}-\mathrm{E}_h\|_{L^2}$		$\|\mathrm{H}-\mathrm{H}_h\|_{L^2}$	
Mesh size	N_{dof}	Error	Order	Error	Order	N_{dof}	Error	Order	Error	Order
1/2	720	2.27e − 1	–	2.35e − 1	–	1,440	3.13e − 2	–	3.36e − 2	–
1/4	5,184	6.02e − 2	1.9	6.68e − 2	1.8	10,368	4.00e − 3	3.0	4.44e − 3	2.9
1/8	39,168	1.54e − 2	2.0	1.78e − 2	1.9	78,336	4.93e − 4	3.0	5.53e − 4	3.0

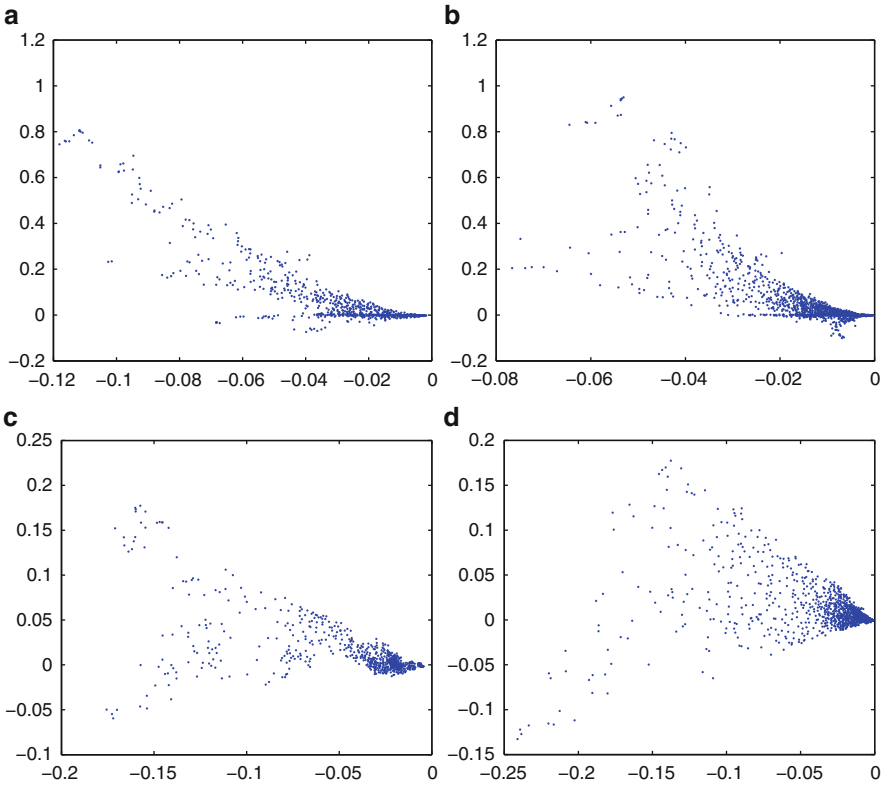


Fig. 1 Eigenvalue distribution of the reduced system. Within parentheses it is (number of eigenvalues located below $y = 0$)/(total number of eigenvalues). (**a**) HDG-$\mathbb{P}_1, \omega = 2\pi$, (241/720). (**b**) HDG-$\mathbb{P}_2, \omega = 2\pi$, (519/1,440). (**c**) HDG-$\mathbb{P}_1, \omega = 8\pi$, (262/720). (**d**) HDG-$\mathbb{P}_2, \omega = 8\pi$, (536/1,440)

5 Conclusion Remarks and Comments on Future Work

We have studied a HDG formulation for the solution of 3D Maxwell's equations and shown the characterization of the reduced system and the numerical traces. We will consider more realistic problems with higher order HDG methods and iterative solvers in a close future.

References

1. Bernardo Cockburn, Jayadeep Gopalakrishnan, and Raytcho Lazarov. Unified hybridization of discontinuous Galerkin, mixed, and continuous Galerkin methods for second order elliptic problems. *SIAM J. Numer. Anal.*, 47(2):1319–1365, 2009.

2. Bernardo Cockburn, Feng-yan Li, and Chi-Wang Shu. Locally divergence-free discontinuous Galerkin methods for the Maxwell equations. *J. Comput. Phys.*, 194(2):588–610, 2004.
3. Victorita Dolean, Hugo Fol, Stéphane Lanteri, and Ronan Perrussel. Solution of the time-harmonic Maxwell equations using discontinuous Galerkin methods. *J. Comput. Appl. Math.*, 218:435–445, 2008.
4. Loula Fezoui, Stéphane Lanteri, Stéphanie Lohrengel, and Serge Piperno. Convergence and stability of a discontinuous Galerkin time-domain method for the 3d heterogeneous Maxwell equations on unstructured meshes. *ESAIM: Math. Model. and Numer. Anal.*, 39(6):1149–1176, 2005.
5. Jan S. Hesthaven and Tim Warburton. Nodal high-order methods on unstructured grids: I. time-domain solution of maxwell's equations. *J. Comput. Phys.*, 181(1):186–221, 2002.
6. Jan S. Hesthaven and Tim Warburton. *Nodal discontinuous Galerkin methods - algorithms, analysis, and applications*. Springer, New York, 2008.
7. Paul Houston, Ilaria Perugia, and Dominik Schötzau. Mixed discontinuous Galerkin appproximation of the Maxwell's equations. *SIAM J. Numer. Anal.*, 42(1):434–459, 2004.
8. Ngoc-Cuong Nguyen, Jaime Peraire, and Bernardo Cockburn. Hybridizable discontinuous Galerkin methods for the time-harmonic Maxwell's equations. *J. Comput. Phys.*, 230(19):7151–7175, 2011.

Locally Implicit Discontinuous Galerkin Methods for Time-Domain Maxwell's Equations

L. Moya

Abstract An attractive feature of discontinuous Galerkin (DG) spatial discretization is the possibility of using locally refined space grids to handle geometrical details. However, when combined with an explicit integration method to numerically solve a time-dependent partial differential equation, this readily leads to unduly large step size restrictions caused by the smallest grid elements. If the local refinement is strongly localized such that the ratio of fine to coarse elements is small, the unduly step size restrictions can be overcome by blending an implicit and an explicit scheme where only solution variables living at fine elements are implicitly treated. The counterpart of this approach is having to solve a linear system per time step. But due to the assumed small fine to coarse elements ratio, the overhead will also be small while the solution can be advanced in time with step sizes determined by the coarse elements.

We propose to present two locally implicit methods for the time-domain Maxwell's equations. Our purpose is to compare the two with DG spatial discretization so that the most efficient one can be advocated for future use. Finally we will present a preliminary numerical investigation to increase the order of convergence.

1 Introduction

We consider time-domain Maxwell's equations

$$\begin{cases} \varepsilon\, \partial_t E &= \text{curl}\, H - \sigma E - J_E, \\ \mu\, \partial_t H &= -\text{curl}\, E, \end{cases} \tag{1}$$

L. Moya (✉)
Inria Sophia Antipolis – Méditerranée, NACHOS project-team, 2004 route des Lucioles, BP 93, 06902 Sophia Antipolis Cedex, France
e-mail: ludovic.moya@inria.fr

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_14,

where E and H denote the electric and magnetic field, J_E is the given source current and ε, μ and σ are coefficients representing dielectric permittivity, magnetic permeability and conductivity. After discretization in space by a DG method we obtain the system

$$\begin{cases} M^{\varepsilon}\,\partial_t E &= SH - DE + M^{\varepsilon} f^E, \\ M^{\mu}\,\partial_t H &= -S^T E + M^{\mu} f^H, \end{cases} \tag{2}$$

where for convenience we use the same notations for the electric and magnetic fields E and H as in the space-continuous case. For more details on DG spatial discretization refer to [2,7]. The matrices M^{ε}, M^{μ} are the DG mass matrices which contain the values of ε and μ. The matrix S emanates from the discretization of the curl operator. The matrix D is associated with the dissipative conduction term $-\sigma E$. The functions f^E and f^H are associated with source terms; f^E represents the given source current $-J_E$, but f^E and f^H may also contain Dirichlet boundary data.

We can give an equivalent formulation of the semi-discrete Maxwell system (2) without mass matrix

$$\begin{cases} \partial_t E &= SH - DE + f^E, \\ \partial_t H &= -S^T E + f^H. \end{cases} \tag{3}$$

It is obtained by a transformation based on the Cholesky decompositions of M^{ε} and M^{μ} and results obtained for (3) applied to (2) and vice versa, see e.g. [1]. For convenience we use the same notations in (2) and (3) and we will proceed with the transformed system (3). S emanates from an appropriate DG discretization for the Maxwell problem under consideration. This means that (refer to [6])

$$S \sim h^{-1}, \quad \text{for } h \to 0, \tag{4}$$

where the parameter h denotes the maximum diameter of the grid elements.

A popular integration method for (3) is the second order Leap-Frog scheme (LF2) that we write in the three-stage form, emanating from Verlet's method, see [7]

$$\begin{cases} \dfrac{H^{n+\frac{1}{2}} - H^n}{\Delta t/2} &= -S^T E^n + f^H(t_n), \\ \dfrac{E^{n+1} - E^n}{\Delta t} &= SH^{n+\frac{1}{2}} - \frac{1}{2} D(E^{n+1} + E^n) + \frac{1}{2}(f^E(t_{n+1}) + f^E(t_n)), \\ \dfrac{H^{n+1} - H^{n+\frac{1}{2}}}{\Delta t/2} &= -S^T E^{n+1} + f^H(t_{n+1}), \end{cases} \tag{5}$$

where $\Delta t = t_{n+1} - t_n$ denotes the step size and upper indices refer to time levels. This method has consistency two, is explicit in S, conditionally stable with a critical time step size proportional to h^{-1}, determined by the smallest grid element, see [1]. Then (5) readily leads to unduly large step size restrictions. An alternative is the second order, unconditionally stable Crank-Nicolson method (CN2) that we write in the three-stage form

$$\begin{cases} \dfrac{H^{n+\frac{1}{2}} - H^n}{\Delta t/2} & = -S^T E^n + f^H(t_n), \\ \dfrac{E^{n+1} - E^n}{\Delta t} & = SH^{n+1} - \frac{1}{2}D(E^{n+1} + E^n) + \frac{1}{2}(f^E(t_{n+1}) + f^E(t_n)), \\ \dfrac{H^{n+1} - H^{n+\frac{1}{2}}}{\Delta t/2} & = -S^T E^{n+1} + f^H(t_{n+1}), \end{cases} \tag{6}$$

which only differ in the middle stage in the time level for H. For consistency and stability we refer to [10]. The expense for the implicit computation is too large to consider (6) as an attractive alternative to (5), especially in 3D (see e.g. [10]).

If the ratio of fine to coarse elements is small the unduly step size restriction of (5) and the overhead of (6) can be overcome by blending the two methods with a locally implicit approach where only variables living at fine elements are implicitly treated.

2 The Locally Implicit Methods

2.1 Component Splitting

The set of grid elements is assumed to be partitioned into two subsets, one made of the smallest elements that will be treated implicitly and the other one of the remaining elements for explicit treatment. In line with this splitting the problem unknowns are reordered as (see [2, 7])

$$E = \begin{pmatrix} E_e \\ E_i \end{pmatrix} \quad \text{and} \quad H = \begin{pmatrix} H_e \\ H_i \end{pmatrix}, \tag{7}$$

where the indices i and e are associated to the elements of the subset treated implicitly and explicitly, respectively. Likewise S, D, f^E and f^H are split as form

$$S = \begin{pmatrix} S_e & -A_{ei} \\ -A_{ie} & S_i \end{pmatrix}, \quad D = \begin{pmatrix} D_e & 0 \\ 0 & D_i \end{pmatrix}, \quad f^E = \begin{pmatrix} f_e^E \\ f_i^E \end{pmatrix}, \quad f^H = \begin{pmatrix} f_e^H \\ f_i^H \end{pmatrix}. \tag{8}$$

For the specific meaning of the block-entries of S refer to [2]. Inserting this splitting into the semi-discrete system (3) we obtain the system of ODEs

$$\begin{cases} \partial_t E_e &= S_e H_e - A_{ei} H_i - D_e E_e + f_e^E(t), \\ \partial_t E_i &= S_i H_i - A_{ie} H_e - D_i E_i + f_i^E(t), \\ \partial_t H_e &= -S_e^T E_e + A_{ie}^T E_i + f_e^H(t), \\ \partial_t H_i &= -S_i^T E_i + A_{ei}^T E_e + f_i^H(t). \end{cases} \tag{9}$$

2.2 The Locally Implicit Method from [2, 7]

This method is a blend of the explicit method (5) and the implicit method (6) applied to (9)

$$\begin{cases} \dfrac{H_e^{n+1/2} - H_e^n}{\Delta t/2} = -S_e^T E_e^n + A_{ie}^T E_i^n + f_e^H(t_n), \\ \dfrac{E_e^{n+1/2} - E_e^n}{\Delta t/2} = S_e H_e^{n+1/2} - A_{ei} H_i^n - D_e E_e^n + f_e^E(t_n), \end{cases}$$

$$\begin{cases} \dfrac{E_i^{n+1} - E_i^n}{\Delta t} = S_i \left(\dfrac{H_i^{n+1} + H_i^n}{2} \right) - A_{ie} H_e^{n+1/2} \\ \qquad\qquad -D_i \left(\dfrac{E_i^{n+1} + E_i^n}{2} \right) + \dfrac{f_i^E(t_{n+1}) + f_i^E(t_n)}{2}, \\ \dfrac{H_i^{n+1} - H_i^n}{\Delta t} = -S_i^T \left(\dfrac{E_i^{n+1} + E_i^n}{2} \right) + A_{ei}^T E_e^{n+1/2} + \dfrac{f_i^H(t_{n+1}) + f_i^H(t_n)}{2}, \end{cases}$$

$$\begin{cases} \dfrac{E_e^{n+1} - E_e^{n+1/2}}{\Delta t/2} = S_e H_e^{n+1/2} - A_{ei} H_i^{n+1} - D_e E_e^{n+1} + f_e^E(t_{n+1}), \\ \dfrac{H_e^{n+1} - H_e^{n+1/2}}{\Delta t/2} = -S_e^T E_e^{n+1} + A_{ie}^T E_i^{n+1} + f_e^H(t_{n+1}). \end{cases} \tag{10}$$

Note that (10) is a symmetric method guaranteeing second-order consistency. Furthermore the implicitness is restricted to E_i and H_i living on the fine grid elements.

2.3 The Locally Implicit Method from [8]

This method is also a blend of (5) and (6) applied to the generic semi-discrete Maxwell system (3)

$$\begin{cases} \dfrac{H^{n+1/2}-H^n}{\Delta t/2} & = -S^T E^n + f^H(t_n), \\ \dfrac{E^{n+1}-E^n}{\Delta t} & = S_0 H^{n+1/2} + \dfrac{1}{2} S_1 (H^n + H^{n+1}) \\ & \quad -\dfrac{1}{2} D(E^n + E^{n+1}) + \dfrac{1}{2}(f^E(t_n) + f^E(t_{n+1})), \\ \dfrac{H^{n+1}-H^{n+1/2}}{\Delta t/2} & = -S^T E^{n+1} + f^H(t_{n+1}), \end{cases} \tag{11}$$

where $S_0 + S_1 = S$ is a general matrix splitting. The method is implicit in S_1 and explicit in S_0. For $S_0 = 0$ we recover (6) and for $S_1 = 0$ method (5). For more details on the definition of the matrix splitting refer to [8].

2.4 Convergence

We are interested in convergence of both methods (10) and (11). Jan Verwer has proven in [8] that the subdivision into coarse and fine elements is not detrimental to the second-order ODE convergence of the method (11), under stable simultaneous space-time grid refinement towards the true underlying PDE solution.

Theorem 1. *Let $f^H(t)$, $f^E(t) \in C^2[0,T]$ and suppose a Lax-Richtmyer stable space-time grid refinement $\Delta t \sim h$, $h \to 0$. On the interval $[0,T]$ the approximations H^n and E^n of method (11) then converge with temporal order two to $H^h(t)$ and $E^h(t)$.*[1]

In [6] we have proven that the component splitting can be detrimental to the convergence order of the method (10) (order reduction).

Theorem 2. *Let $f^H(t)$, $f^E(t) \in C^2[0,T]$ and suppose a Lax-Richtmyer stable space-time grid refinement $\Delta t \sim h$, $h \to 0$. On $[0,T]$ the approximations H_e^n, H_i^n, E_e^n and E_i^n of method (10) then converge to $H_e^h(t)$, $H_i^h(t)$, $E_e^h(t)$ and $E_i^h(t)$*

(i) At least at first order,

(ii) At least at second order, if in addition $A_{ei}^T S_e H_e^{h(1)}(t) = \mathcal{O}(\Delta t^{-1})$ for $h \to 0$.

[1] $H^h(t)$ and $E^h(t)$ denote the exact solutions of the Maxwell problem under consideration, restricted to the space grid.

We can only guarantee the first-order convergence of method (10) and we have put forward a sufficient condition on the true solution of the PDE problem for second-order.

3 Numerical Results

We solve the two-dimensional (2D) Transverse Magnetic (TM) model for the components $E^z(x, y, t)$, $H^x(x, y, t)$ and $H^y(x, y, t)$

$$\begin{cases} \mu \dfrac{\partial H^x}{\partial t} = -\dfrac{\partial E^z}{\partial y}, \\ \mu \dfrac{\partial H^y}{\partial t} = \dfrac{\partial E^z}{\partial x}, \\ \varepsilon \dfrac{\partial E^z}{\partial t} = \dfrac{\partial H^y}{\partial x} - \dfrac{\partial H^x}{\partial y} - \sigma E^z - J_E^z. \end{cases} \tag{12}$$

In the following, we set $\varepsilon = \mu = 1$ and $\sigma = 0$. Equation (12) are space discretized using a DG method formulated on quadrangular meshes. In the implementation of this DG method, the approximation of the electromagnetic field components within a quadrangle relies on a nodal $\mathbb{Q}_2$ interpolation method in order that the spatial error will not be detrimental to the temporal convergence order in the PDE sense.[2] For the numerical experiments the integration interval in time is $[0, 1]$ and we use uniform meshes[3] (see Fig. 1). For the definition of the numerical critical step size we refer to [6]. Finally, for the implicit treatment we choose the region $[0.4, 0.6] \times [0.4, 0.6]$.

We consider the propagation of an eigenmode in a unitary PEC cavity. In this problem there is no source term ($J_E^z = 0$) and the exact solution is given by

$$\begin{cases} H^x(x, y, t) = -\dfrac{k}{\sqrt{k^2 + l^2}} \sin(l\pi x) \cos(k\pi y) \sin(\sqrt{k^2 + l^2}\pi t), \\ H^y(x, y, t) = \dfrac{l}{\sqrt{k^2 + l^2}} \cos(l\pi x) \sin(k\pi y) \sin(\sqrt{k^2 + l^2}\pi t), \\ E^z(x, y, t) = \sin(l\pi x) \sin(k\pi y) \cos(\sqrt{k^2 + l^2}\pi t). \end{cases} \tag{13}$$

[2] The a priori convergence analysis for this DGTD method based on a centered numerical flux, formulated on simplicial meshes, provides a convergence rate in $\mathcal{O}(h^p)$ for a p-th interpolation order.

[3] We focus on the convergence order and not on the practical virtues of locally implicit methods. In the latter case it would be more appropriate to consider non-conforming meshes with a local refinement.

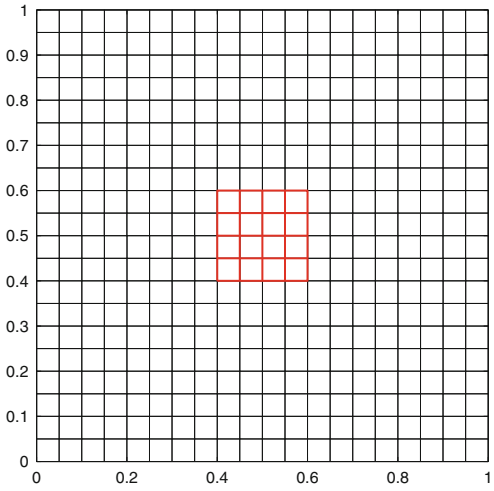


Fig. 1 Example of an uniform quadrangular mesh considered in the numerical tests: # elements = 400, # degrees of freedom (DOF) = 3, 600 for the DGTD-$\mathbb{Q}_2$ method, *red region* = implicit treatment

Table 1 Max. L^2-norm error and convergence of methods (10) and (11) based on the DGTD-$\mathbb{Q}_2$ method

	Max. error in L^2-norm	
# DOF	Method (10)	Method (11)
900	0.0012	5.6029e − 4
3,600	5.4236e − 4	1.3603e − 4
8,100	3.5759e − 4	6.0265e − 5
14,400	2.7164e − 4	3.3821e − 5
Convergence rate	1.0762	2.0254

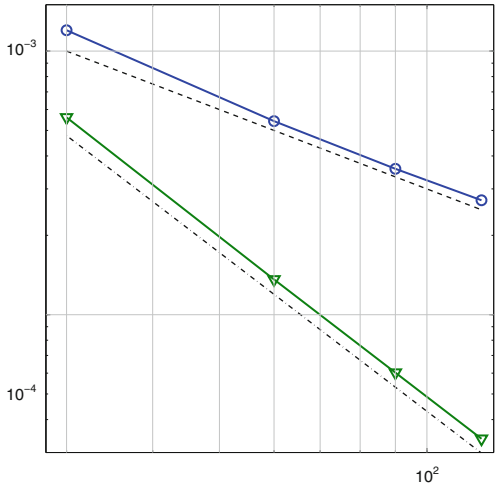


Fig. 2 Loglog convergence plots (max. L^2-norm error as function of the square root of # DOF) for methods (10) and (11) based on the DGTD-$\mathbb{Q}_2$ method. ∘-*marker* and ∇-*marker* represent the results obtained for (10) and (11) methods, respectively; *the dashed* and *dotted/dashed lines* have slope one and two for first and second-order convergence, respectively

For numerical tests we put $k = l = 1$. To estimate the convergence order we measure the maximal L^2-norm of the error for different meshes of increased resolution. The sufficient condition of Theorem 2 is not satisfied and the results given in Table 1 and Fig. 2 clearly confirm the theoretical behavior i.e. the first-order convergence of (10) and the second-order for (11).

Table 2 Max. L^2-norm error and convergence of CO, LEX and GEX based on the second-order method (11) and a DG-$\mathbb{Q}_4$ spatial discretization

	Max. error in L^2-norm		
# DOF	CO	LEX	GEX
2,500	1.4038e − 7	1.3865e − 7	6.6628e − 8
10,000	2.2363e − 8	5.9528e − 8	4.7422e − 9
22,500	8.0648e − 9	2.1371e − 8	1.2795e − 9
40,000	3.9302e − 9	1.0434e − 8	5.3285e − 10
Convergence rate	2.5812	1.8574	3.4971

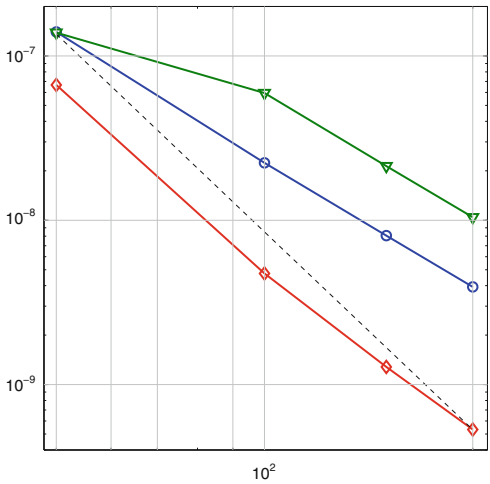


Fig. 3 Loglog convergence plots for DGTD-$\mathbb{Q}_4$ methods. ∘-*marker*, ∇-*marker* and ◇-*marker* represent the results obtained for CO, LEX and GEX methods, respectively; *the dashed line* has slope four for fourth-order convergence

4 Fourth-Order Methods (Preliminary Investigation)

We consider three well-known techniques in ODE setting to increase the order of convergence at fourth-order: symmetric composition, local and global Richardson extrapolation denoted CO, LEX and GEX, respectively. For details on composition methods we refer to [3, 9], on Richardson extrapolation to [1, 4, 5]. Note that the composition methods with orders beyond two are restricted to problems with small (non-stiff) dissipative terms [1]. We use (11) as the second-order basis method and we consider the numerical test of Sect. 3 with a nodal $\mathbb{Q}_4$ interpolation method. The results given in Table 2 and Fig. 3 show an order reduction for each method, especially for LEX and CO, due to the subdivision into coarse and fine elements.[4] However, the accuracy of these latter methods will certainly be very interesting.

[4] We have also conduct the numerical test for the fully explicit case, i.e. with (5) as the basis method, and there was no reduction order for each fourth-order technique and this particular test case (no source term, no dissipative term).

Acknowledgements The author wishes to acknowledge the many and important contributions to this work by Stéphane Descombes, Stéphane Lanteri and Jan Verwer.

References

1. Botchev, M.A., Verwer, J.G.: Numerical Integration of Damped Maxwell Equations. SIAM J. Sci. Comput. **31**, 1322–1346 (2009)
2. Dolean, V., Fahs, H., Fezoui, L., Lanteri, S.: Locally implicit discontinuous Galerkin method for time domain electromagnetics. J. Comput. Phys. **229**, 512–526 (2010)
3. Hairer, E.,Lubich, C., Wanner, G.: Geometric Numerical Integration. Second edition, Springer-Verlag, Berlin (2002)
4. Hairer, E., Wanner, G.:Solving Ordinary Differential Equations II - Stiff and Differential-Algebraic problems. Second edition, Springer-Verlag, Berlin (1996)
5. Kulikov, G.Yu.: Local theory of extrapolation methods. Numer. Algor. **53**, 321–342 (2010)
6. Moya, L., Verwer, J.G.: Temporal convergence of a locally implicit discontinuous Galerkin method for Maxwell's equations. (2011) http://hal.inria.fr/inria-00565217/fr/
7. Piperno, S.: Symplectic local time-stepping in non-dissipative DGTD methods applied to wave propagation problem. ESAIM: M2AN **40**, 815–841 (2006)
8. Verwer, J.G.: Component splitting for semi-discrete Maxwell equations. BIT Numer. Math. (2010) doi: 10.1007/s10543-010-0296-y
9. Verwer, J.G.: Composition methods, Maxwell's and source term. (2010) http://oai.cwi.nl/oai/asset/17036/17036A.pdf
10. Verwer, J.G., Botchev, M.A.: Unconditionaly stable integration of Maxwell's equations. Linear Algebra and its Applications **431**, 300–317 (2009)

Application of the Level-Set Method to a Mixed-Mode and Curvature Driven Stefan Problem

D. den Ouden, F.J. Vermolen, L. Zhao, C. Vuik, and J. Sietsma

Abstract This study focuses on the dissolution and growth of small possibly initially non-smooth particles within a diffusive phase. The dissolution or growth of the particle is assumed to be affected by concentration gradients of a single chemical element within the diffusive phase at the particle boundary caused by diffusion and by an interface reaction. The combined formulation results in a mixed-mode formulation. The moving boundary problem is solved using a level-set method and finite-element techniques such as SUPG. The appropriate meshes are derived using a fixed background mesh and the level-set function. We experimentally show that these techniques give mass-conserving solutions in the limit of infinite resolution, give a linear experimental order of convergence, can handle arbitrary particles and give the possibility to incorporate surface tensions using the Gibbs-Thomson effect and the local curvature.

1 Introduction

Metalworking of alloys is a complex process that involves several physical phenomena, such as dislocation movement, grain recrystallisation and secondary phase precipitation, that influence the applicability of the object. The influence of these phenomena has been studied and documented by processes based on

D. den Ouden (✉) · L. Zhao
Materials innovation institute, Mekelweg 2, 2628, CD Delft, The Netherlands
e-mail: d.denouden@m2i.nl; Lie.Zhao@tudelft.nl

F.J. Vermolen · C. Vuik
Delft University of Technology, DIAM, Mekelweg 4, 2628 CD Delft, The Netherlands
e-mail: F.J.vermolen@tudelft.nl; c.vuik@tudelft.nl

J. Sietsma
Delft University of Technology, MS&E, Mekelweg 2, 2628, CD Delft, The Netherlands
e-mail: j.sietsma@tudelft.nl

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_15,

"trial and error". An analytical approach to investigate these aspects could verify the obtained experimental results and hence deliver relevant understanding of the physical behaviour of alloys during metalworking.

During the last two decades various models for the nucleation and growth of precipitates in alloys have been proposed and evaluated. Several models can be found in [2, 3, 7, 9, 11, 13]. Most of these models are based on the assumption of a spherical shape for the precipitates, thereby neglecting the occurrence of other precipitate shapes. Our study aims at modelling the growth and dissolution of arbitrary particles shapes and under various physical conditions.

In this paper we will first give a brief overview of the moving boundary model used for the description of growth and dissolution. Subsequently, we discuss two important numerical aspects of our proposed method and finally we discuss the results obtained with our method with regard to convergence and the dependence of the model on the curvature and the mixed-mode boundary conditions.

2 The Model

In this section we describe the models used to simulate the dissolution and growth of a particle in a diffusive phase. We will employ the level-set method [6] for the description of the boundary, whereas we let dissolution and growth be influenced by diffusive and reactional fluxes.

2.1 The Stefan Problem

This research uses a modification of the original Stefan problem defined by Jožef Stefan in 1890 (see [1]). We consider a particle $\Omega_P(t)$ embedded in a diffusive phase $\Omega_D(t)$ and denote the interface between particle and diffusive phase by $\Gamma(t)$, i.e. the moving boundary. Within the particle $\Omega_P(t)$, we assume a fixed concentration c_p and within the diffusive phase $\Omega_D(t)$ the concentration $c(\mathbf{x}, t)$ is modelled using the standard diffusion equation

$$\frac{\partial c}{\partial t}(\mathbf{x}, t) = D\Delta c(\mathbf{x}, t), \qquad \text{for } \mathbf{x} \in \Omega_D(t), t > 0, \tag{1}$$

where D is the diffusion constant. At the outer boundary of Ω_D, i.e. $\partial\Omega_D(t) \setminus \Gamma(t)$, we assume a no flux condition.

At the interface $\Gamma(t)$, we assume that two phenomena occur, the first being a first-order reaction in which the crystal structure of the particle phase $\Omega_P(t)$ is transformed into the crystal structure of the diffusive phase $\Omega_D(t)$ or vice versa, the second being the flux of atoms away from or to the interface $\Gamma(t)$ into the bulk diffusive phase $\Omega_D(t)$. The net flux at the interface can therefore be described by the boundary condition

$$K\left(c_s(\mathbf{x},t)-c(\mathbf{x},t)\right)=D\frac{\partial c}{\partial n}(\mathbf{x},t)+c(\mathbf{x},t)v_n(\mathbf{x},t),\quad \text{for } \mathbf{x}\in\Gamma(t),\ t>0, \tag{2}$$

where $v_n(\mathbf{x},t)$ denotes the velocity of the interface point $\mathbf{x}\in\Gamma(t)$ in the direction of the outward normal $\mathbf{n}$. The parameter K represents the speed with which atoms can transfer between phases. The problem definition is completed by the standard Stefan condition

$$c_p v_n(\mathbf{x},t)=D\frac{\partial c}{\partial n}(\mathbf{x},t)+c(\mathbf{x},t)v_n(\mathbf{x},t),\qquad \text{for } \mathbf{x}\in\Gamma(t),\ t>0. \tag{3}$$

The solubility of the considered element at the particle boundary inside the diffusive phase $c_s(\mathbf{x},t)$ is modelled using the Gibbs-Thomson effect [8]

$$c_s(\mathbf{x},t)=c_s^{\infty}(t)\exp\left(\zeta\kappa(\mathbf{x},t)\right), \tag{4}$$

where $c_s^{\infty}(t)$ is the solubility of the considered element equilibrium in the diffusive phase, ζ a factor composed of the surface energy, the molar volume of the particle, the gas constant and the temperature in Kelvin, and $\kappa(\mathbf{x},t)$ the local interface curvature at the interface point $\mathbf{x}\in\Gamma(t)$.

2.2 Level Set Method

To capture the motion and location of the interface $\Gamma(t)$, described fully by Eqs. (1)–(4), we employ the level-set method as first introduced in [6]. To this end, we define the signed-distance function $\phi(\mathbf{x},t)$ by

$$\phi(\mathbf{x},t)=\begin{cases}+\min\limits_{\mathbf{y}\in\Gamma(t)}\|\mathbf{y}-\mathbf{x}\|_2 \text{ if } \mathbf{x}\in\Omega_P(t),\\ -\min\limits_{\mathbf{y}\in\Gamma(t)}\|\mathbf{y}-\mathbf{x}\|_2 \text{ if } \mathbf{x}\in\Omega_D(t).\end{cases} \tag{5}$$

In $\Omega=\Omega_P(t)\cup\Omega_D(t)$, the normal $\mathbf{n}(\mathbf{x},t)$ and the curvature $\kappa(\mathbf{x},t)$ can be directly calculated from the level-set function $\phi(\mathbf{x},t)$ in a standard way.

Following [5], we let $\phi(\mathbf{x},t)$ evolve according to the convection equation

$$\frac{\partial\phi}{\partial t}(\mathbf{x},t)+v_n^{ex}(\mathbf{x},t)\,\|\nabla\phi(\mathbf{x},t)\|_2=0, \tag{6}$$

where the value of v_n^{ex} is obtained by solving the Laplace equation at each time t for v_n^{ex} in Ω with the Dirichlet boundary condition $v_n^{ex}(\mathbf{x},t)=v_n(\mathbf{x},t)$ for all $x\in\Gamma(t)$ and a homogeneous Neumann boundary condition at $\partial\Omega\setminus\Gamma(t)$.

To ensure that $\phi(\mathbf{x},t)$ remains a signed-distance function at all steps of the simulation, we employ the reinitialisation technique introduced in [12].

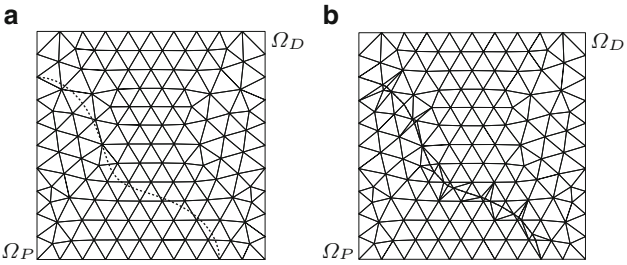


Fig. 1 Examples of the tessellations $\mathscr{T}$ and $\mathscr{T}_C$, where the *dashed line* represents the interface

3 Numerical Methods

To obtain the solution for the system given in the previous section, we apply the SUPG finite-element technique [14] with an Implicit Euler time integration for Eq. (1) and an Explicit Euler time integration for Eq. (6). In this paper, we will divide the domain $\Omega_D \cup \Omega_P$ into triangles, resulting into the background tessellation $\mathscr{T}$. We will not solve any of the equations on this tessellation but on different tessellations, which are all determined from the background tessellation $\mathscr{T}$ and the interface curve.

3.1 Tessellations

Given the finite-element approximation $\boldsymbol{\phi}^n$ on the tessellation $\mathscr{T}$, we interpolate over the edges of $\mathscr{T}$ to find the discrete points that constitute $\Gamma(t^n)$. If the distance between an original vertex and a new boundary point is less then δ times the corresponding edge length, we move the vertex to the nearest boundary point. With the resulting set of edge points and boundary points, we will construct a new tessellation $\mathscr{T}_C(t^n)$ with the constraint that there is a continuous curve of edges through the boundary points. An example of a tessellation $\mathscr{T}$ and the resulting tessellation $\mathscr{T}_C(t^n)$ can be found in Fig. 1. Afterwards a tessellation $\mathscr{T}_D(t^n)$ of $\Omega_D(t^n)$ is extracted by taking all triangles and line elements where $\boldsymbol{\phi}^n \leq 0$, on which the weak formulation of Eq. (1) can be solved using a finite-element method with linear basis functions.

3.2 Material Derivative

Derivation of the weak formulation of Eq. (1), application of the finite-element method and subsequent application of Implicit Euler time integration will result

in a formulation in which integrals over the elements from $\mathscr{T}_D(t^{n+1})$ and $\Gamma(t^{n+1})$ appear using $\mathbf{c}^n$, the concentration at the previous time-step. As $\mathbf{c}^n$ is only defined on the tessellation $\mathscr{T}_D(t^n)$, we employ a technique derived from the Moving Mesh Algorithm proposed in [10]. We replace the partial time derivative in Eq. (1) with a material derivative, where the mesh velocity at any point $\mathbf{x}_i \in \mathscr{T}_D(t^{n+1})$ is defined by

$$\mathbf{v}_m(\mathbf{x}_i, t^{n+1}) = (\mathbf{x}_i - \mathbf{y}_i) \,/\, \left(t^{n+1} - t^n\right), \tag{7}$$

with the point $\mathbf{y}_i \in \mathscr{T}_D(t^n)$ such that $\|\mathbf{x}_i - \mathbf{y}_i\|_2$ is minimal.

4 Results

4.1 The Experimental Accuracy: Dissolution of Planar and Circular Particles

To investigate the accuracy and rate of convergence of the current method, we simulate the dissolution of a planar particle on $[0, 1]\times[0, 1]$ with initial width 0.615. We will simulate on a regular square mesh with $N + 1$ points in both coordinate directions. We take $c(\mathbf{x}, 0) = 0.3$, $c_p = 0.45$, $D = 1$, $K = 10^3$, $\zeta = 0$, $\delta = 0.3$ and let c_s^∞ increase linearly from 0.301 to 0.33 between $t = 0$ and $t = 0.1$ and 0.33 for all other t. The value $K = 10^3$ implies that $c(\mathbf{x}, t) \to c_s(t)$ for $\mathbf{x} \in \Gamma(t)$ and $t \to \infty$. We simulate until $t = 3$ and we take as time-step Δt the minimum of the CFL condition with $\texttt{CFL} = 0.25$ and $10h^2/D$ with h the characteristic spacing of the tessellation $\mathscr{T}_C$. The first $N/2$ time-steps CFL will be divided by 10, due to the initial discontinuities at the interface. With these physical and numerical parameters we also simulated the dissolution of a circular particle on the unit circle with initial radius 0.615. We simulated on circular meshes with approximately $2N^2$ elements.

The relative errors obtained at $t = 3$ with respect to the initial mass of the system and the equilibrium width/radius of the particle are shown in Table 1. Further, the normalized standard deviation of the results around the mean width/radius of the interface is shown. This normalisation results from division of the standard deviation by the mean width/radius of the particle, hence this statistical parameter acts as an indicator of maintaining the topology of the particle. From these results we can see that we have a linear experimental order of convergence in mass conservation, particle size and topology preservation.

4.2 Simulation of Particle Growth

To investigate the performance of our method for the growth of precipitates, we simulate the growth of an initially diamond shaped particle with the four vertexes

Table 1 The relative errors achieved at $t = 3$ for the dissolution of the planar particle and the circular particle

	Planar particle			Circular particle		
N	Mass error	Width error	Topology error	Mass error	Radius error	Topology error
16	2.8333×10^{-3}	1.7847×10^{-2}	1.7234×10^{-4}	1.2751×10^{-2}	6.4919×10^{-2}	7.3113×10^{-3}
32	5.2828×10^{-4}	3.3279×10^{-3}	1.0093×10^{-4}	5.0091×10^{-3}	2.7712×10^{-2}	2.8867×10^{-3}
64	1.9233×10^{-4}	1.2077×10^{-3}	4.7998×10^{-5}	2.2274×10^{-3}	1.3184×10^{-2}	1.0972×10^{-3}
128	1.7299×10^{-5}	1.0556×10^{-4}	2.9833×10^{-5}	1.2196×10^{-3}	7.4270×10^{-3}	6.4786×10^{-4}

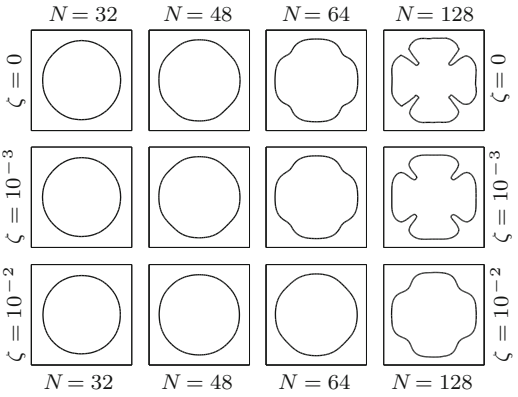


Fig. 2 The particle shape at $t = 0.2$ for the various mesh sizes N and values for the interface parameter ζ for the growth of an initially diamond shaped particle on $[-0.5, 0.5]\times[-0.5, 0.5]$

located at $(\pm0.1, \pm0.1)$ within the computational domain $[-0.5, 0.5]\times[-0.5, 0.5]$. We choose the physical parameters as $c_0 = 0.3, c_s^\infty = 0.1, c_p = 0.45, D = 1, K = 10^3$. We use the same type of meshes and numerical parameters as for the planar dissolution and simulate until $t = 0.2$. Due to the occurrence of oscillations at the interface, of which the number is expected to increase for finer meshes, we expect that these oscillations will be amplified, consistent with the Mullins-Sekerka instability [4]. Figure 2 shows the particle shape at $t = 0.2$ for various mesh sizes N and various values for the interface parameter ζ.

If we compare the horizontally sequential figures, we see with the increase of the number of grid points an increase in the loss of topology, which is manifested by the occurrence of simultaneously more and less local growth of the particle. This coincides with the analytical results obtained in [4] and the increase in the number of oscillations at the interface with an increase in the number of grid points.

If we compare the vertically sequential figures, we see with the increase of the surface parameter ζ a decrease in the loss of topology, which is manifested by the more rounded shape of the particle at higher ζ. This coincides with the physical effect of the Gibbs-Thomson effect [8], which represents the minimisation of free energy by reducing the overall curvature of the interface. As a result the most favourable shape of the particle at equilibrium for a given system is spherical under incorporation of surface tensions, which is also apparent from the simulations.

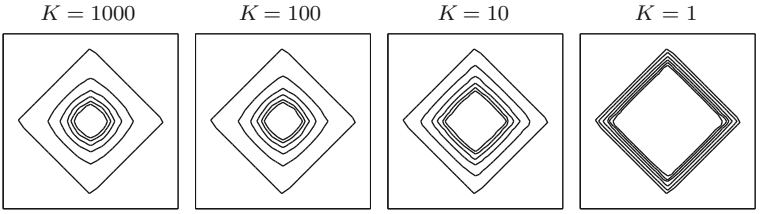


Fig. 3 The particle shape at six equally spaced times between $t = 0$ and $t = 1$ for the various values for the interface reaction speed K for the dissolution of an initially diamond shaped particle on $[-0.5, 0.5] \times [-0.5, 0.5]$

4.3 Simulation of Mixed-Mode Dissolution

In the previous examples we chose the diffusivity D and the interface reaction speed K such that the growth and dissolution are mainly limited by the diffusion of the solute towards or away from the interface. To investigate the influence of the parameter K, we simulate the dissolution of an initially diamond shaped particle with the four vertexes located at $(\pm 0.42, \pm 0.42)$ within the computational domain $[-0.5, 0.5] \times [-0.5, 0.5]$. We choose the physical parameters as $c_0 = 0.3, c_p = 0.45, D = 1, \zeta = 0$ and let c_s^∞ increase linearly from 0.301 to 0.35 between $t = 0$ and $t = 0.1$ and 0.35 for all other t. We use the same type of meshes and numerical parameters as for the planar dissolution with $N = 32$ and simulate until $t = 1$. We test for the values $K = 1{,}000$, $K = 100$, $K = 10$ and $K = 1$. The results for these four simulations can be found in Fig. 3.

The reduction of K from 1,000 to 1 has two distinct effects. First the speed at which the particle dissolves becomes more constant over time. Furthermore the shape of the particles stays more like the initial shape for lower values of K. Both of these effects are caused by the direct balance of the diffusive flux and the interface-reaction flux. A high K will cause the diffusive effect to become rate-determining, with as a result that at a position on the interface where a high diffusive flux is possible, i.e. at the corners, a high dissolution rate is obtained. This causes the particle to round off. If a lower value of K is assumed, then the reactional flux becomes limiting, which is determined by the concentration at the interface, leading to a more constant dissolution rate along the interface. Our tests also show that for high grid resolutions our method correctly captures the movement of sharp cornered particles under various physical settings.

5 Conclusions

We numerically solve a modified Stefan problem for the dissolution and growth of particles. The movement of the interface between the particle and diffusive phase is determined by long-distance diffusion and interfacial reactions on the moving

boundary. The numerical method is based on a level-set method to track the moving interface. All off the PDEs are solved using a SUPG finite-element method [14]. To cope with the moving boundary in an accurate way, the triangulation around the interface is adjusted by introducing additional nodes on the element sides that are intersected by the moving interface. By experiment, we show that the current method is mass-conserving and converges linearly with respect to increasing grid resolution. Furthermore, we show that the method gives consistent results for all modes of diffusion, interface reactions and surface tensions.

Acknowledgements This research was carried out under the project number M41.5.09341 in the framework of the Research Program of the Materials innovation institute M2i (www.m2i.nl).

References

1. J. Crank, *Free and moving boundary problems* (Clarendon Press, Oxford, 1984)
2. A. Deschamps, Y. Brechet, Acta Mater. **47**
3. R. Kampmann, R. Wagner, *Materials Science and Technology – A Comprehensive Treatment*, vol. 5 (VCH, Weinheim, 1991)
4. W.W. Mullins, R.F. Sekerka, J. Appl. Phys. **34**
5. S.J. Osher, R.P. Fedkiw, R.P., *Level Set Methods and Dynamic Implicit Surfaces* (Springer, Netherlands, 2002)
6. S.J. Osher, J.A. Sethian, J. Comput. Phys. **79**
7. D. den Ouden, F.J. Vermolen, L. Zhao, C. Vuik, J. Sietsma, Comp. Mater. Sci. **50**
8. D.A. Porter, K.E. Easterling, *Phase Transformations in Metals and Alloys*, 2nd edn. (Chapmann & Hall, London, 1992)
9. J.D. Robson, Mater. Sci. Technol. **20**
10. A. Segal, C. Vuik, F.J. Vermolen, J. Comp. Phys. **141**
11. F. Soisson, A. Barbu, G. Martin, Acta Mater. **44**
12. M. Sussman, P. Smereka, S. Osher, J. Comput. Phys. **114**
13. F.J. Vermolen, E. Javierre, C. Vuik, L. Zhao, S. van der Zwaag, Comp. Mater. Sci. **39**
14. X. Xing, P. Wei, M.Y. Wang, Int. J. Numer. Meth. Eng. **82**

On an Efficient Family of Simultaneous Methods for Finding Polynomial Multiple Zeros

J. Džunić, M.S. Petković, and L.D. Petković

Abstract An iterative method for the simultaneous determination of multiple zeros of algebraic polynomials is stated. This method is more efficient compared to all existing simultaneous methods based on fixed point relations. To attain very high computational efficiency, a suitable correction resulting from Li-Liao-Cheng's two-point fourth-order method of low computational complexity is applied. The presented convergence analysis shows that the convergence rate of the basic method is increased from three to six using this special type of correction and applying only ν additional polynomial evaluations per iteration, where ν is the number of distinct zeros. Computational aspects and some numerical examples are given to demonstrate high computational efficiency and very fast convergence of the proposed method.

1 Introduction

The aim of this paper is to construct an iterative method for the simultaneous determination of all multiple zeros of a polynomial with a very high computational efficiency. Actually, the proposed method is ranked as the most efficient among existing methods in the class of simultaneous methods for approximating polynomial multiple zeros based on fixed point relations. The presented iterative formula relies on the fixed point relation of Gargantini's type [3]. A very high computational

J. Džunić (✉) · M.S. Petković
Faculty of Electronic Engineering, Department of Mathematics, University of Niš, 18000, Niš, Serbia
e-mail: jovana.dzunic@elfak.ni.ac.rs; msp@junis.ni.ac.rs

L.D. Petković
Faculty of Mechanical Engineering, Department of Mathematics, University of Niš, 18000, Niš, Serbia
e-mail: ljiljana@masfak.ni.ac.rs

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_16,

efficiency is attained by employing suitable corrections which enable very fast convergence (equal to six) with minimal additional computational costs. In fact, these corrections arise from Li-Liao-Cheng's two-point root-solver [4] with optimal order of convergence four. More details about multi-point methods may be found, e.g., in [8] and [10].

The paper is organized as follows. In Sect. 2 we present the improved iterative method for the simultaneous determination of polynomial multiple zeros, starting from a suitable fixed-point relation. The convergence theorem stated in Sect. 3 asserts that the convergence order of the proposed method is six. Finally, Sect. 4 contains an analysis of computational efficiency which shows that the proposed simultaneous method is the most efficient among all existing methods based on fixed point relations. In addition, two numerical examples are given to demonstrate exceptional convergence speed of the proposed method.

2 Accelerated Simultaneous Method

Let $f(z) = \prod_{j=1}^{\nu}(z-\zeta_j)^{\mu_j}$ be a monic polynomial of degree n with multiple real or complex zeros $\zeta_1, \ldots, \zeta_\nu$ of respective multiplicities $\mu_1, \ldots, \mu_\nu$ ($\nu \le n$), and let

$$u(z) = \frac{f(z)}{f'(z)} = \Big[\frac{d}{dz}\log f(z)\Big]^{-1} = \Big(\sum_{j=1}^{\nu}\frac{\mu_j}{z-\zeta_j}\Big)^{-1}. \tag{1}$$

To construct an iterative method for the simultaneous determination of polynomial multiple zeros, we single out the term $z-\zeta$ from (1) and derive the following fixed point relation

$$\zeta_i = z - \frac{\mu_i}{\frac{1}{u(z)} - \sum\limits_{\substack{j\in \boldsymbol{I}_\nu \\ j\ne i}} \frac{\mu_j}{z-\zeta_j}} \quad (i \in \boldsymbol{I}_\nu := \{1,\ldots,\nu\}). \tag{2}$$

This relation was used in [3] for the construction of iterative methods for the simultaneous inclusion of multiple zeros of polynomials in complex circular arithmetic.

Let $z_1, \ldots, z_\nu$ be distinct approximations to the zeros $\zeta_1, \ldots, \zeta_\nu$. Setting $z = z_i$ and substituting the zeros ζ_j by some approximations z_j^* in the right-hand side of (2), one obtains the following iterative method

$$\hat{z}_i = z_i - \frac{\mu_i}{\frac{1}{u(z_i)} - \sum\limits_{\substack{j\in \boldsymbol{I}_\nu \\ j\ne i}} \frac{\mu_j}{z_i - z_j^*}} \quad (i \in \boldsymbol{I}_\nu) \tag{3}$$

for the simultaneous determination of all multiple zeros of the polynomial f. Here $\hat{z}_i$ denotes a new approximation to the zero ζ_i. The choice $z_j^* = z_j$ in (3) gives the third-order method of Ehrlich-Aberth's type for multiple zeros

$$\hat{z}_i = z_i - \frac{\mu_i}{\dfrac{1}{u(z_i)} - \displaystyle\sum_{\substack{j\in \boldsymbol{I}_\nu \\ j\neq i}} \dfrac{\mu_j}{z_i - z_j}} \quad (i \in \boldsymbol{I}_\nu), \tag{4}$$

see [1, 2]. Furthermore, putting Schröder's approximations $z_j^* = z_j - \mu_j u(z_j)$ in (3), the following accelerated method of the fourth order is obtained (see [5]),

$$\hat{z}_i = z_i - \frac{\mu_i}{\dfrac{1}{u(z_i)} - \displaystyle\sum_{\substack{j\in \boldsymbol{I}_\nu \\ j\neq i}} \dfrac{\mu_j}{z_i - z_j + \mu_j u(z_j)}} \quad (i \in \boldsymbol{I}_\nu). \tag{5}$$

Note that the iterative method (5) reduces to Nourein's method [6] in the case of simple zeros.

Regarding (2)–(5), it is evident that the better approximations z_j^* give the more accurate approximations $\hat{z}_i$. Indeed, if $z_j^* \to \zeta_j$ then the right-hand side of (3) tends to ζ_i. We apply this idea to construct a higher order method.

The iterative method (5) of the fourth order is obtained using Schröder's method $z_j^* = z_j - \mu_j u(z_j)$ of the second order. Further acceleration of the convergence speed can be obtained by using methods of higher order for finding a single multiple zero. In this paper we use the following two-point method for solving nonlinear equations proposed in [4]

$$\hat{z} = z - u(z) \cdot \frac{\beta + \gamma t(z)}{1 + \delta t(z)}, \quad t(z) = \frac{f'(z - \theta u(z))}{f'(z)}, \tag{6}$$

where

$$\theta = \frac{2m}{m+2}, \quad \beta = -\frac{m^2}{2}, \quad \gamma = \frac{m(m-2)}{2}\Big(\frac{m}{m+2}\Big)^{-m}, \quad \delta = -\Big(\frac{m}{m+2}\Big)^{-m}$$

and m is the multiplicity of the wanted zero ζ of a function f (not necessarily algebraic polynomial in general). The order of convergence of the iterative method (6) is four, that is,

$$\hat{z} - \zeta = O_M\big((z-\zeta)^4\big) \tag{7}$$

holds (for the proof, see [4]). Here O_M is a symbol which points to the fact that two complex numbers w_1 and w_2 have moduli of the same order (that is, $|w_1| = O(|w_2|)$, O is the Landau symbol), written as $w_1 = O_M(w_2)$.

In the sequel, we substitute z by the approximation z_j of ζ_j and m by the corresponding multiplicity μ_j of ζ_j. The approximation z_j^* appearing in (3) is

calculated by (6), that is,

$$z_j^* = z_j - u_j \cdot \frac{\beta_j + \gamma_j t_j}{1 + \delta_j t_j},$$

where we put $u_j = u(z_j)$, $t_j = f'(z_j - \theta_j u_j)/f'(z_j)$ and

$$\theta_j = \frac{2\mu_j}{\mu_j + 2},\ \beta_j = -\frac{\mu_j^2}{2},\ \gamma_j = \frac{\mu_j(\mu_j - 2)}{2}\Big(\frac{\mu_j}{\mu_j + 2}\Big)^{-\mu_j},\ \delta_j = -\Big(\frac{\mu_j}{\mu_j + 2}\Big)^{-\mu_j}.$$

Now, from (3) we obtain a new method for the simultaneous approximation of all simple or multiple zeros of a given polynomial,

$$\hat{z}_i = z_i - \frac{\mu_i}{\dfrac{1}{u(z_i)} - \displaystyle\sum_{\substack{j \in \boldsymbol{I}_v \\ j \neq i}} \dfrac{\mu_j}{z_i - z_j + u_j \cdot \frac{\beta_j + \gamma_j t_j}{1 + \delta_j t_j}}} \quad (i \in \boldsymbol{I}_v). \tag{8}$$

3 Convergence Theorem

Theorem 1. *If initial approximations $z_1, \ldots, z_\nu$ are sufficiently close to the distinct zeros $\zeta_1, \ldots, \zeta_\nu$ of a given polynomial, then the order of convergence of the simultaneous method (8) is six.*

Proof. Let us introduce the errors of approximations $\varepsilon_j = z_j - \zeta_j$, $\hat{\varepsilon}_j = \hat{z}_j - \zeta_j$. According to the conditions of Theorem 1, we can assume that $\varepsilon_i = O_M(\varepsilon_j)$ for any pair $i, j \in \boldsymbol{I}_v$. Let $\varepsilon \in \{\varepsilon_1, \ldots, \varepsilon_n\}$ be the error of maximal modulus with $\varepsilon_j = O_M(\varepsilon)$ $(j \in \boldsymbol{I}_v)$.

For brevity, let

$$z_j^* = z_j - u_j \cdot \frac{\beta_j + \gamma_j t_j}{1 + \delta_j t_j}, \quad d_i = \sum_{\substack{j \in \boldsymbol{I}_v \\ j \neq i}} \frac{\mu_j (z_j^* - \zeta_j)}{(z_i - \zeta_j)(z_i - z_j^*)}.$$

Then, starting from (8) and using (1) we obtain

$$\hat{z}_i = z_i - \frac{\mu_i}{\dfrac{\mu_i}{\varepsilon_i} + \displaystyle\sum_{\substack{j \in \boldsymbol{I}_v \\ j \neq i}} \frac{\mu_j}{z_i - \zeta_j} - \sum_{\substack{j \in \boldsymbol{I}_v \\ j \neq i}} \frac{\mu_j}{z_i - z_j^*}} = z_i - \frac{\mu_i \varepsilon_i}{\mu_i - \varepsilon_i d_i}, \tag{9}$$

and hence

$$\hat{\varepsilon}_i = \hat{z}_i - \zeta_i = \varepsilon_i - \frac{\mu_i \varepsilon_i}{\mu_i - \varepsilon_i d_i} = \frac{-\varepsilon_i^2 d_i}{\mu_i - \varepsilon_i d_i}. \tag{10}$$

According to (7) we have $d_i = O_M(\varepsilon^4)$ and from (10) we find

$$\hat{\varepsilon} = O_M(\varepsilon^6),$$

since the denominator of (10) tends to μ_i when $\varepsilon_i \to 0$. Therefore, the order of convergence of the simultaneous method (8) is six. □

4 Computational Aspects

From a practical point of view, it is of great importance to estimate the computational efficiency of any iterative root-finding method since it is closely connected to the features such as the number of necessary numerical operations in computing the zeros with the required accuracy, the convergence speed, processor time of a computer, etc. The knowledge of the computational efficiency is of particular interest in designing a package of root-solvers. More details about this topic may be found in [7, Chap. 6].

In this section we compare the convergence behavior and computational efficiency of the methods (4), (5) and the new simultaneous method (8). This comparison procedure is entirely justified since the analysis of efficiency given in [7, Chap. 6] for several computing machines showed that the method (5) has the highest computational efficiency in the class of simultaneous methods based on fixed point relations.

Comparing the iterative formulas (5) and (8) we observe that the new formula (8) requires ν new polynomial evaluations per iterations in relation to (5). Hence we conclude that the minimal computational efficiency of the iterative method (8) appears when $\nu = n$, that is, when all zeros are simple. For this reason we will consider this "worst case" in our computational analysis. In a similar way as in [9] and several other papers in the topic, we estimated computational efficiency of the iterative methods (4), (5) and (8) using the *efficiency index* given by

$$E(\mathrm{IM}) = \frac{\log r}{d}, \tag{11}$$

where r is the order of convergence of the iterative method (IM), and d is its computational cost. The computation cost d is evaluated using the total number of basic arithmetic operations per iteration taken with certain *weights* depending on the execution times of operations, see [9] for details.

We calculated the percent ratios

$$\rho_{8,4}(n) = (E((8), n)/E((4), n) - 1) \cdot 100 \ \ (\text{in } \%), \qquad \text{(F/EA \%)}$$

$$\rho_{8,5}(n) = (E((8), n)/E((5), n) - 1) \cdot 100 \ \ (\text{ in } \%), \qquad \text{(F/N\%)}$$

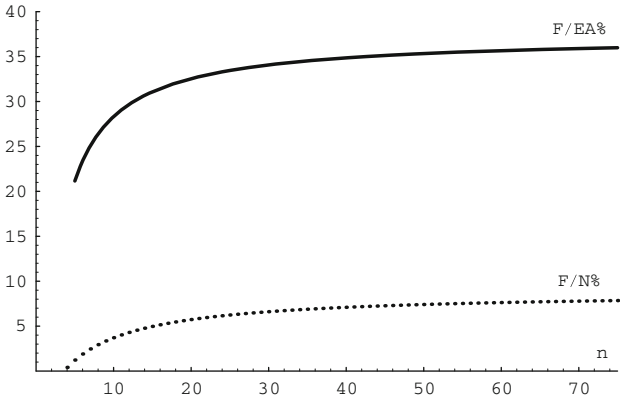


Fig. 1 Ratios of efficiency indices

where EA, N and F stand for the method (4) of Ehrlich-Aberth's type, the method (5) of Nourein's type and the new method (8), respectively. These ratios are graphically displayed in Fig. 1 as the functions of the polynomial degree n and show the (percent) improvement of computational efficiency of the new method (8) in relation to the methods (4) and (5). In Fig. 1 $\rho_{8,4}(n)$ is drawn by full line and $\rho_{8,5}(n)$ by dotted line.

It is evident from Fig. 1 that the new method (8) is more efficient than the methods (4) and (5). The improvement is especially expressive in regard to the method (4) of Ehrlich-Aberth's type (F/EA % – full line). Having in mind the mentioned fact on the dominant efficiency of the Nourein-like method, it follows that the proposed family of simultaneous methods (8) is the *most efficient method* for the simultaneous determination of polynomial multiple zeros in the class of methods based on fixed point relations.

To demonstrate the convergence behavior of the methods (4), (5) and (8), we tested a number of polynomial equations; for illustration, among a number of tested algebraic polynomials we selected two examples. To present the results of the third iteration, we applied the computational software package *Mathematica* with multiple-precision arithmetic.

As a measure of accuracy of the obtained approximations, we calculated Euclid's norm

$$e^{(m)} := ||\boldsymbol{z}^{(m)} - \boldsymbol{\zeta}||_2 = \left(\sum_{i=1}^{\nu} \left| z_i^{(m)} - \zeta_i \right|^2 \right)^{1/2} \qquad (m = 0, 1, \ldots),$$

where $\boldsymbol{z}^{(m)} = \left(z_1^{(m)}, \ldots, z_\nu^{(m)}\right)$ and $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_\nu)$.

Table 1 Euclid's norm of the errors – Example 1

Methods →	(4)	(5)	(8)
$e^{(1)}$	2.81(−1)	1.62(−1)	1.80(−1)
$e^{(2)}$	2.61(−3)	6.00(−5)	9.03(−7)
$e^{(3)}$	2.93(−9)	1.92(−18)	1.21(−39)

Table 2 Euclid's norm of the errors – Example 2

Methods →	(4)	(5)	(8)
$e^{(1)}$	1.10(−1)	5.57(−2)	2.18(−2)
$e^{(2)}$	7.24(−5)	2.38(−7)	7.44(−13)
$e^{(3)}$	1.64(−14)	3.34(−29)	3.65(−75)

Example 1. Methods (4), (5) and (8) were applied for the simultaneous approximation to the zeros of the polynomial

$$f_{18}(z) = (z+1)^2(z+2)^3(z^2-2z+2)^2(z^2+1)^2(z-2)^3(z+2-i)^2.$$

The following starting approximations were selected ($e^{(0)} \approx 1.50$)

$$z_1^{(0)} = -1.3 + 0.2i,\ z_2^{(0)} = -2.2 - 0.3i,\ z_3^{(0)} = 1.3 + 1.2i,\ z_4^{(0)} = 0.7 - 1.2i,$$
$$z_5^{(0)} = -0.2 + 0.8i,\ z_6^{(0)} = 0.2 - 1.3i,\quad z_7^{(0)} = 2.2 - 0.3i,\ z_8^{(0)} = -2.2 + 0.7i.$$

The entries of the maximal errors obtained in the first three iterations are given in Table 1.

Example 2. In order to find the zeros of the polynomial

$$f_{20}(z) = (z+1)^2(z+3)^3(z^2-2z+2)^2(z-1)^3(z^2-4z+5)^2(z^2+4z+5)^2,$$

we applied the same methods. The starting approximations were ($e^{(0)} \approx 1.43$)

$$z_1^{(0)} = -1.3 + 0.2i,\ z_2^{(0)} = -2.8 - 0.2i,\ z_3^{(0)} = 1.2 + 1.3i,$$
$$z_4^{(0)} = 0.8 - 1.2i,\quad z_5^{(0)} = 0.8 - 0.3i,\quad z_6^{(0)} = -1.8 + 1.2i,$$
$$z_7^{(0)} = -1.8 - 1.2i,\ z_8^{(0)} = 1.8 + 0.8i,\quad z_9^{(0)} = 1.8 - 1.2i.$$

The entries of the maximal errors obtained in the first three iterations are given in Table 2.

From Tables 1 and 2 and a number of tested polynomial equations we can conclude that the proposed family (8) produces approximations of considerably great accuracy; two iterative steps are usually sufficient in solving most practical problems when initial approximations are reasonably close to the zeros.

The presented analysis of computational efficiency shows that the family (8) is more efficient than all existing methods for multiple zeros based on fixed point relations.

Acknowledgements This work was supported by the Serbian Ministry of Science under grant 174022.

References

1. O. Aberth, Iteration methods for finding all zeros of a polynomial simultaneously, Math. Comp. **27** (1973), 339–344.
2. L.W. Ehrlich, A modified Newton method for polynomials, Comm. ACM **10** (1967), 107–108.
3. I. Gargantini, Further application of circular arithmetic: Schröder-like algorithms with error bound for finding zeros of polynomials, SIAM J. Numer. Anal. **15** (1978), 497–510.
4. S. Li, X. Liao, L. Cheng, A new fourth-order iterative method for finding multiple roots of nonlinear equations, Appl. Math. Comput. **215** (2009), 1288–1292.
5. G. V. Milovanović, M. S. Petković, On the convergence order of a modified method for simultaneous finding polynomial zeros, Computing **30** (1983) 171–178.
6. A. W. M. Nourein, An improvement on two iteration methods for simultaneously determination of the zeros of a polynomial, Internat. J. Comput. Math. **6** (1977), 241–252.
7. M. S. Petković, *Iterative Methods for Simultaneous Inclusion of Polynomial Zeros*, (Springer-Verlag, Berlin-Heidelberg-New York, 1989), pp. 221–249.
8. M. S. Petković, On a general class of multipoint root-finding methods of high computational efficiency, SIAM J. Numer. Anal. **47** (2010), 4402–4414.
9. M. S. Petković, M. R. Milošević, D. M. Milošević, New higher-order methods for the simultaneous inclusion of polynomial zeros, Numer. Algorithms, **58** (2011), 179–201.
10. M. S. Petković, L. D. Petković, Families of optimal multipoint methods for solving nonlinear equations: a survey, Appl. Anal. Discrete Math. **4** (2010), 1–22.

Multilevel Sparse Kernel-Based Interpolation Using Conditionally Positive Definite Radial Basis Functions

E.H. Georgoulis, J. Levesley, and F. Subhan

Abstract A multilevel sparse kernel-based interpolation (MLSKI) method, suitable for moderately high-dimensional function interpolation problems has been recently proposed in (Georgoulis et al. Multilevel sparse kernel-based interpolation, submitted for publication). The method uses both level-wise and direction-wise multilevel decomposition of structured or mildly unstructured interpolation data sites in conjunction with the application of kernel-based interpolants with different scaling in each direction. The multilevel interpolation algorithm is based on a hierarchical decomposition of the data sites, whereby at each level the detail is added to the interpolant by interpolating the resulting residual of the previous level. On each level, anisotropic radial basis functions (RBFs) are used for solving a number of small interpolation problems, which are subsequently linearly combined to produce the interpolant. Here, we investigate the use of conditionally positive definite RBFs within the MLSKI setting, thus extending the results from (Georgoulis et al. Multilevel sparse kernel-based interpolation, submitted for publication), where (strictly) positive definite RBFs are used only.

1 Introduction

Over the last four decades, radial basis functions (RBFs) have been successfully applied to (scattered) data interpolation/approximation in $\mathbb{R}^d \times \mathbb{R}$ (see, e.g., [29] and the references therein for a literature review). The interest on kernel-based and, in particular, on RBF interpolants can be traced in their ability to produce global interpolants of user-defined smoothness without the shortcomings of multivariate polynomial interpolation. These interpolants admit generally good convergence

E.H. Georgoulis (✉) · J. Levesley · F. Subhan
Department of Mathematics, University of Leicester, University Road, Leicester, LE1 7RH, UK
e-mail: Emmanuil.Georgoulis@le.ac.uk; j.levesley@le.ac.uk; f.subhan@hotmail.com

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_17,

properties and they can be implemented in (essentially) dimension-independent fashion, making them potentially attractive for a number of applications.

Despite the above attractive properties, RBF interpolation can be cumbersome in practice. Solving the resulting linear system is challenging due to both the density and the ill-conditioning of the resulting interpolation matrix (see, e.g., [22]). A number of techniques have been proposed to deal with the ill-conditioning of the interpolation system [3,8,9,11–13]. The introduction of RBFs with compact support [28, 30] aims to address the density issue of the interpolation matrix. Moreover, a number of techniques have been developed to reduce the complexity of calculating the interpolant, involving multipole type expansion for a variety of RBFs [3].

Recently, the multilevel kernel-based interpolation (MLSKI) method has been put forward [15,26] as a stable, reduced complexity algorithm for RBF interpolation of moderately high-dimensional problems, i.e., problems with $d = 2, 3, 4$. The numerical results in [15] suggest that the respective MLSKI algorithm would yield good interpolation results for $d = 5$ or, possibly, higher d also. MLSKI can be viewed as an extension of d-boolean interpolation [7, 24, 27] to kernel-based functions, which, in turn, is closely related to ideas in sparse grid [4, 16, 18, 31] and hyperbolic crosses [1, 23, 25] literature. To accelerate convergence, a hierarchical multilevel framework is used [10, 17, 20, 21].

In [15, 26] it was shown that, regardless of the kernel used, the SKI/MLSKI algorithm is able to solve interpolation problems on sparse grids up to 114,690 centers for $d = 3$, and up to 331,780 centers for $d = 4$, *without* taking advantage of the possibility of parallel implementation of SKI or MLSKI.

The discussion in [15] is confined to strictly positive definite kernels. In this short contribution, we investigate the applicability of the MLSKI algorithm to conditionally positive definite RBFs, such as the multiquadric (MQ) $\varphi(r) = \sqrt{c^2 + r^2}$ and the thin plate spline (TPS2) $\varphi(r) = r^2 \log r$. The MLSKI algorithm is tested in practice for a number of relevant test cases for $d = 2, 3, 4$. Numerical experiments with these conditionally positive definite RBFs, presented below, suggest that the new algorithm is also numerically stable and efficient for the reconstruction of large data in $\mathbb{R}^d \times \mathbb{R}$, for $d = 2, 3, 4$, with tens or even hundreds of thousands of data points.

2 Multilevel Sparse Kernel-Based Interpolation

For simplicity, let $\Omega := [0, 1]^d$, and consider $u : \Omega \to \mathbb{R}$. For a multi-index $\mathbf{l} = (l_1, \ldots, l_d) \in \mathbb{N}^d$, we define the family of directionally uniform grids $\{\mathbb{X}_\mathbf{l} : \mathbf{l} \in \mathbb{N}^d\}$, in Ω, with meshsize $h_\mathbf{l} = 2^{-\mathbf{l}} := (2^{-l_1}, \ldots, 2^{-l_d})$. That is, $\mathbb{X}_\mathbf{l}$ consists of the points $\mathbf{x}_{\mathbf{l},\mathbf{i}} := (x_{l_1,i_1}, \ldots, x_{l_d,i_d})$, with $x_{l_j,i_j} = i_j 2^{-l_j}$, for $i_j = 0, 1, \ldots, 2^{l_j}$, $j = 1, \ldots, d$. The number of nodes $N^\mathbf{l}$ in $\mathbb{X}_\mathbf{l}$ is given by

$$N^\mathbf{l} = \prod_{i=1}^{d} (2^{l_i} + 1).$$

If $h_{l_i} = 2^{-n}$, for all $i = 1, \cdots, d$, $\mathbb{X}_{\mathbf{l}}$ is the uniform *full grid* of level n, having size $N = (2^n + 1)^d$; this will be denoted by $\mathbb{X}^{n,d}$. We also consider the following subset of $\mathbb{X}^{n,d}$,

$$\tilde{\mathbb{X}}^{n,d} := \bigcup_{|\mathbf{l}|_1 = n+(d-1)} \mathbb{X}_{\mathbf{l}}, \tag{1}$$

with $|\mathbf{l}|_1 := l_1 + \cdots + l_d$, which will be referred to as the *sparse grid of level n in d dimensions*.

We want to evaluate the interpolant at the constituent sub-grids $\mathbb{X}_{\mathbf{l}}$. As the constituent grids admit different density in each coordinate direction, we shall make use the anisotropic RBFs (cf. [15]). To this end, for each multi-index $\mathbf{l} = (l_1, \ldots, l_d)$, we define the transformation matrix $A_{\mathbf{l}} \in \mathbb{R}^{d\times d}$ by

$$A_{\mathbf{l}} := \operatorname{diag}(2^{l_1}, \ldots, 2^{l_d}).$$

For each $\mathbf{x} \in \Omega$, the anisotropic RBF interpolant $S_{A_{\mathbf{l}}}$ of u at the points of $\mathbb{X}_{\mathbf{l}}$ is then defined by

$$S_{A_{\mathbf{l}}}(\mathbf{x}) := \sum_{j=1}^{N^{\mathbf{l}}} c_j \varphi(\|A_{\mathbf{l}}(\mathbf{x} - \mathbf{x}_j)\|) + \sum_{k=1}^{M} d_k p_k(A_{\mathbf{l}}\mathbf{x}), \tag{2}$$

for a conditionally positive definite kernel φ of order m; $\{p_1, \ldots, p_M\}$ is a basis of the polynomial space π^d_{m-1}. The coefficients $c_j \in \mathbb{R}$ and $d_k \in \mathbb{R}$ are chosen so that the interpolation conditions

$$S_{A_{\mathbf{l}}}|_{\mathbb{X}_{\mathbf{l}}} = u|_{\mathbb{X}_{\mathbf{l}}},$$

and the constraints

$$\sum_{j=1}^{N^{\mathbf{l}}} c_j\, p_l(\mathbf{x}_j) = 0, \qquad \text{for } 1 \le l \le M,$$

are simultaneously satisfied (see [2, 5, 6], for more on anisotropic RBFs). The wellposedness of the interpolation problem is guaranteed [2] as a direct consequence of the invertibility of the scaling matrix A and the π^d_{m-1}-unisolvency of the data sites $\{\mathbf{x}_1, \ldots, \mathbf{x}_{N^I}\}$

To construct the *sparse kernel-based interpolant* (SKI, for short) $\tilde{S}_n$ on the sparse grid $\tilde{\mathbb{X}}^{n,d}$, the sub-grid interpolants $S_{A_{\mathbf{l}}}$ are linearly combined using the formula

$$\tilde{S}_n(\mathbf{x}) = \sum_{q=0}^{d-1} (-1)^q \binom{cd-1}{q} \sum_{|\mathbf{l}|_1 = n+(d-1)-q} S_{A_{\mathbf{l}}}(\mathbf{x}). \tag{3}$$

This formula has been used in the context of d-boolean lagrange polynomial interpolation [7], and in the sparse-grid combination technique for the numerical solution of elliptic partial differential equations using the finite element method [14, 16].

The sparse kernel-based interpolant $\tilde{S}_n$ can be implemented in a quite straightforward fashion by utilising existing, fast, RBF interpolation algorithms: the only modification needed is the introduction of a scaling for each sub-grid problem. We note that each interpolation problem can be solved completely independently, rendering the resulting SKI method ideally suited for implementation in parallel computers. Moreover, there are substantial savings in the computational complexity as d grows (cf. [15], for a detailed discussion).

The basic SKI method, described above is used within a multilevel interpolation algorithm. This is possible, by observing that the sparse grids from lower to higher level are nested.

The *multilevel SKI* (MLSKI, for short) algorithm is initialised by computing the SKI $\tilde{S}_{n_0}$ at the coarsest designated sparse grid $\tilde{\mathbb{X}}^{n_0,d}$ and set $\Delta_0 := \tilde{S}_{n_0}$. Then, for $k = 1, \ldots n$, we compute Δ_k to be the sparse grid interpolant of the residual $u - \sum_{j=0}^{k-1} \Delta_j$ on $\tilde{\mathbb{X}}^{k,d}$. The resulting multilevel sparse kernel based interpolant is then given by

$$\tilde{S}_n^{\mathrm{ML}} := \sum_{j=0}^{n} \Delta_j .$$

3 Numerical Experiments

The numerical experiments presented below have been implemented in MATLAB© on a 3.16 GHz CPU with 3.24 GB of RAM. No attempt has been made to use fast algorithms for the constituent RBF interpolation problems. The CPU-times presented below are approximate and are included as a conceptual measure of run-time complexity. We compare the SKI and MLSKI algorithms with a standard RBF interpolation method on full grids and with its, standard, multilevel version on full grids (i.e., a multilevel implementation of the RBF interpolation algorithm [10]), henceforth denoted by MLRBF.

Let u_{F3D} be a three-dimensional version of Franke's function u_{F3D}, given by

$$\begin{aligned} u_{F3D}(x_1, x_2, x_3) = & \frac{3}{4} e^{(-(9x_1-2)^2-(9x_2-2)^2-(9x_3-2)^2)/4} \\ & + \frac{3}{4} e^{-((9x_1+1)^2)/49-((9x_2+1)^2)/10-((9x_3+1)^2)/29} \\ & + \frac{1}{2} e^{-((9x_1-7)^2)/4-(9x_2-3)^2-((9x_3-5)^2)/2} \\ & - \frac{1}{5} e^{-((9x_1-4)^2)/4-(9x_2-7)^2-((9x_3-5)^2)} . \end{aligned} \tag{4}$$

In Fig. 1, the root mean-square interpolation error curves for u_{F3D}, using RBF, MLRBF, SKI and MLSKI with MQ basis functions are plotted against the number of data sites N and against the computational time (in seconds). For RBF and

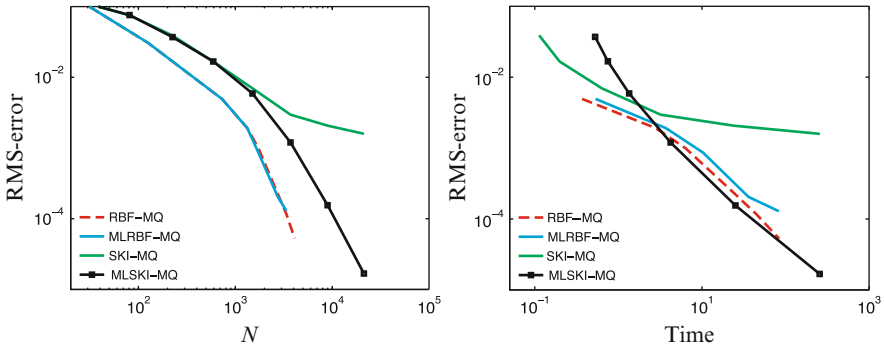


Fig. 1 Convergence of MQ for u_{F3D}. Error evaluated at 125,000 Halton points

MLRBF, N denotes the number of data sites on the full grid $\mathbb{X}^n$, whereas for SKI and MLSKI, N refers to the number of sparse grid nodes (i.e., N = SGnode). We refer to [15, 26] for a discussion on the choice of the shape parameter c. Here, we only note that c is chosen so as to have comparable, safe condition numbers (i.e., smaller than 10^{10}) for all four methods. We observe that RBF/MLRBF interpolation appears to perform better than SKI/MLSKI when the error is plotted against N. The SKI/MLSKI algorithms are, on the other hand, able to calculate larger problems and they do so efficiently in terms of complexity, at least for large N. This is manifested in the RMS-error versus computation time plot in Fig. 1.

Next, we turn our attention to the problem of interpolating five-dimensional data ($d = 4$). We compare the four interpolation methods on a four-variate version of the Franke's function $u_{F4D} : [0, 1]^4 \to \mathbb{R}$, with

$$\begin{aligned} u_{F4D}(x_1, \ldots, x_4) := & \frac{3}{4} e^{(-(9x_1-2)^2-(9x_2-2)^2-(9x_3-2)^2)/4-(9x_4-2)^2)/8} \\ & + \frac{3}{4} e^{-((9x_1+1)^2)/49-((9x_2+1)^2)/10-((9x_3+1)^2)/29-((9x_4+1)^2)/39} \\ & + \frac{1}{2} e^{-((9x_1-7)^2)/4-(9x_2-3)^2-((9x_3-5)^2)/2-((9x_4-5)^2)/4} \\ & - \frac{1}{5} e^{-((9x_1-4)^2)/4-(9x_2-7)^2-((9x_3-5)^2)-((9x_4-5)^2)}. \end{aligned}$$

The convergence history, given in Fig. 2, indicates the ability of MLSKI to compute very accurate interpolants, with substantial savings in computational time, compared to classical RBF/MLRBF.

Next, we repeat the experiment of interpolation of u_{F3D} and u_{F4D}. using the thin-plate spline (TPS2) RBF; the convergence history is given in Figs. 3 and 4, respectively. The choice in the shape parameter is as resulting to safe conditions numbers for all methods. Interestingly, the SKI method seems to converge very slowly, while the MLSKI algorithm appears to perform well.

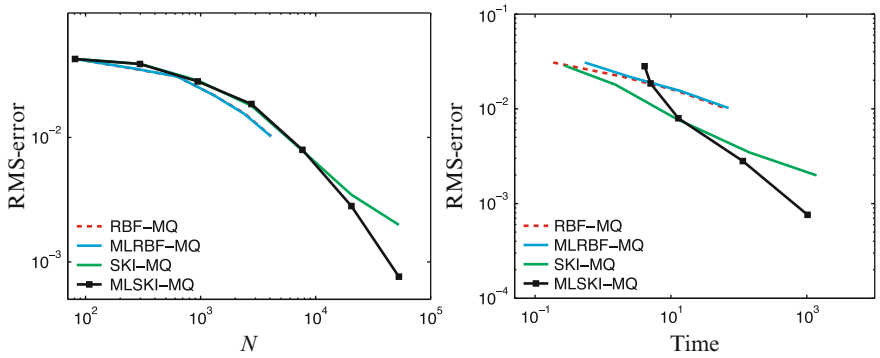


Fig. 2 Convergence of MQ for u_{F4D}. Error evaluated at 194,481 Halton points

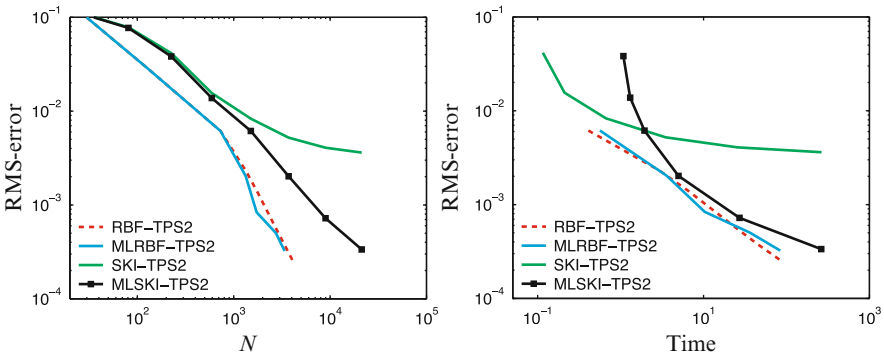


Fig. 3 Convergence of $TPS2$ for u_{F3D}. Error evaluated at 125,000 Halton points

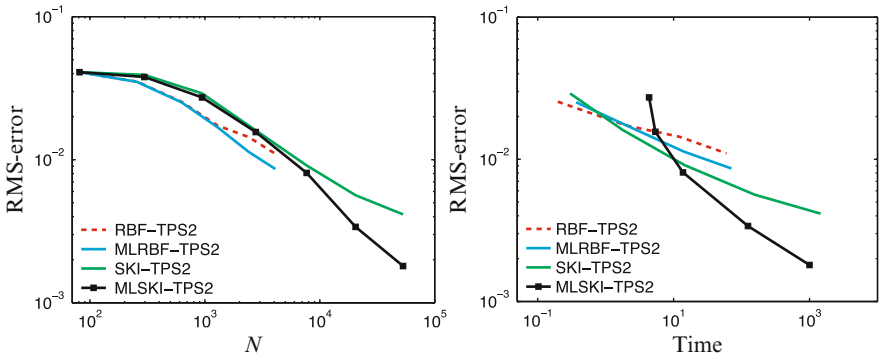


Fig. 4 Convergence of $TPS2$ for u_{F4D}. Error evaluated at 194,481 Halton points

4 Concluding Remarks

The multilevel kernel-based interpolation method, suitable for moderately high-dimensional interpolation problems on carefully structured grids proposed in [15] has been extended to some conditionally positive definite kernels. Comparing the results of the present work and of [15], we observe qualitatively similar convergence behaviour with the MLSKI method being the most efficient as the number of data points grows. Potential advantages of the use of conditionally positive definite kernels over, say Gaussians, are the better understood error behaviour and analysis and the possibility of using RBFs with compact support. We note, however, that our numerical experiments indicate that the parameter choice for which compactly supported RBFs are competitive results to comparable supports with the domain itself.

Currently, the choice of data sets (sparse grids) is highly structured. Some preliminary numerical experiments for SKI on mildly perturbed sparse grids, presented in [26], indicate that the SKI method converges with the same rate, albeit with a somewhat larger constant. Perhaps a more robust methodology for extending the applicability of SKI/MLSKI methods to scattered data is the pre-computation of the values on the corresponding sparse grid data sites via local interpolation. The extension of the SKI/MLSKI method to more general geometries could possibly be handled either by introducing fictitious gridded data sites with suitable data values, taking into account the nature of the data, or by conformally mapping the computational domain [19].

References

1. K. I. Babenko. Approximation by trigonometric polynomials in a certain class of periodic functions of several variables. *Soviet Math. Dokl.*, 1:672–675, 1960.
2. R. Beatson, O. Davydov, and J. Levesley. Error bounds for anisotropic RBF interpolation. *J. Approx. Theory*, 162(3):512–527, 2010.
3. R. K. Beatson, J. B. Cherrie, and C. T. Mouat. Fast fitting of radial basis functions: methods based on preconditioned GMRES iteration. *Adv. Comput. Math.*, 11(2–3):253–270, 1999. Radial basis functions and their applications.
4. H.-J. Bungartz, M. Griebel, and U. Rüde. Extrapolation, combination, and sparse grid techniques for elliptic boundary value problems. *Comput. Methods Appl. Mech. Engrg.*, 116(1–4):243–252, 1994. ICOSAHOM'92 (Montpellier, 1992).
5. G. Casciola, D. Lazzaro, L. B. Montefusco, and S. Morigi. Shape preserving surface reconstruction using locally anisotropic radial basis function interpolants. *Comput. Math. Appl.*, 51(8):1185–1198, 2006.
6. G. Casciola, L. B. Montefusco, and S. Morigi. The regularizing properties of anisotropic radial basis functions. *Appl. Math. Comput.*, 190(2):1050–1062, 2007.
7. F.-J. Delvos. d-variate Boolean interpolation. *J. Approx. Theory*, 34:99–114, 1982.
8. G. E. Fasshauer and Mccourt M. J. Stable evaluation of Gaussian RBF interpolants. *Submitted for publication*, 2011.
9. A. C. Faul and M. J. D. Powell. Proof of convergence of an iterative technique for thin plate spline interpolation in two dimensions. *Adv. Comput. Math.*, 11:183–192, 1999.

10. M. S. Floater and A. Iske. Multistep scattered data interpolation using compactly supported radial basis functions. *J. Comput. Appl. Math.*, 73(1–2):65–78, 1996.
11. B. Fornberg, E. Larsson, and N. Flyer. Stable computation with Gaussian radial basis functions. *SIAM J. Sci. Comput.*, 33(2):869–892, 2011.
12. B. Fornberg and C. Piret. A stable algorithm for flat radial basis functions on a sphere. *SIAM J. Sci. Comput.*, 30(1):60–80, 2007/08.
13. B. Fornberg and G. Wright. Stable computation of multiquadric interpolants for all values of the shape parameter. *Comput. Math. Appl.*, 48(5–6):853–867, 2004.
14. J. Garcke and M. Griebel. On the parallelization of the sparse grid approach for data mining. In S. Margenov, J. Wasniewski, and P. Yalamov, editors, *Large-Scale Scientific Computations, Third International Conference, LSSC 2001, Sozopol, Bulgaria*, volume 2179 of *Lecture Notes in Computer Science*, pages 22–32. Springer, 2001. also as SFB 256 Preprint 721, Universität Bonn, 2001.
15. E. H. Georgoulis, J. Levesley, and F. Subhan. Multilevel sparse kernel-based interpolation. *submitted for publication*.
16. M. Griebel, M. Schneider, and C. Zenger. A combination technique for the solution of sparse grid problems. In *Iterative methods in linear algebra (Brussels, 1991)*, pages 263–281. North-Holland, Amsterdam, 1992.
17. S. J. Hales and J. Levesley. Error estimates for multilevel approximation using polyharmonic splines. *Numer. Algorithms*, 30(1):1–10, 2002.
18. Markus Hegland, Jochen Garcke, and Vivien Challis. The combination technique and some generalisations. *Linear Algebra Appl.*, 420(2–3):249–275, 2007.
19. A. R. H. HERYUDONO AND T. A. DRISCOLL, *Radial basis function interpolation on irregular domain through conformal transplantation*, J. Sci. Comput., 44 (2010), pp. 286–300.
20. A. Iske. Hierarchical scattered data filtering for multilevel interpolation schemes. In *Mathematical methods for curves and surfaces (Oslo, 2000)*, Innov. Appl. Math., pages 211–221. Vanderbilt Univ. Press, Nashville, TN, 2001.
21. A. Iske and J. Levesley. Multilevel scattered data approximation by adaptive domain decomposition. *Numer. Algorithms*, 39(1–3):187–198, 2005.
22. R. Schaback. Error estimates and condition numbers for radial basis function interpolation. *Adv. Comput. Math.*, 3(3):251–264, 1995.
23. A. Schreiber. *The method of Smolyak in multivariate interpolation*. PhD thesis, der Mathematisch-Naturwissenschaftlichen Fakultäten, der Georg-August-Universität zu Göttingen, 2000.
24. W. Sickel and F. Sprengel. Interpolation on sparse grids and tensor products of Nikol′skij-Besov spaces. *J. Comput. Anal. Appl.*, 1(3):263–288, 1999. Dedicated to Professor Paul L. Butzer on the occasion of his 70th birthday.
25. S. A. Smolyak. Quadrature and interpolation of formulas for tensor product of certian classes of functions. *Soviet Math. Dokl.*, 4:240–243, 1963.
26. F. Subhan. Multilevel sparse kernel-based interpolation. *Ph.D. Thesis, University of Leicester*, 2011.
27. V. N. Temlyakov. Approximation of functions with bounded mixed derivative. In *Proc. Steklov Institute Math, 1989.* AMS, 1989.
28. H. Wendland. Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Adv. Comput. Math.*, 4:389–396, 1995.
29. H. Wendland. *Scattered data approximation*, volume 17 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, Cambridge, 2005.
30. Z. Wu. Compactly supported positive definite radial functions. *Adv. Comput. Math.*, 4:283–292, 1995.
31. C. Zenger. Sparse grids. In *Parallel algorithms for partial differential equations (Kiel, 1990)*, volume 31 of *Notes Numer. Fluid Mech.*, pages 241–251. Vieweg, Braunschweig, 1991.

A Numerical Remark on the Time Discretization of Contact Problems in Nonlinear Elasticity

C. Groß, R. Krause, and V. Poletti

Abstract The time discretization of contact-problems in elasticity is a difficult task, since the non-penetration condition at the contact interface can lead to instabilities in displacements, stresses, and energy. For the case of linear elasticity, in (Deuflhard et al., Int J Numer Methods Eng 73(9):1274–1290, 2008), a contact stabilized Newmark scheme has been proposed, which employs a discrete L^2-projection at the contact boundary for stabilization and which can shown to be energy dissipative. Here, we combine this contact-stabilization with an approach presented on (Gonzalez, Comput Methods Appl Mech Eng 190(13–14):1763–1783, 2000) for the time discretization of unconstrained problems in nonlinear mechanics. We apply the resulting combined scheme to contact problems with non-linear non-penetration constraints and non-linear material laws and numerically investigate its behavior. Although our combined scheme is not proven to be energy dissipative, it does not show any decrease in energy and the resulting displacements and forces at the contact boundary show a highly stable behaviour.

1 Introduction

The numerical simulation of elastodynamical contact problems is a challenging task for a variety of reasons. First of all, the contact area is unknown in advance. Secondly, due to the non-penetration constraint, the problem is inherently non-smooth, both in space and in time. In fact, after discretization in space and time, impact and detachment of single points (e.g. nodes of a finite element mesh) of the contact boundary will generally occur during a timestep and not at discrete integration points. This introduces numerical artifacts as oscillation of velocities

C. Groß (✉) · R. Krause · V. Poletti
Institute of Computational Science, University of Lugano, via Giuseppe Buffi 13, CH-6900, Lugano, Switzerland
e-mail: christian.gross@usi.ch; rolf.krause@usi.ch; valentina.poletti@usi.ch

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_18,

or displacements at the contact boundary. Also, the conservation of structural properties like energy might be problematic, see, e.g. [5]. In spite of these difficulties various approaches have been studied, attempting to minimize these problems and to obtain results that are as close as possible to the physical realm, cf. [7].

The widely utilized family of Newmark integrators, while being momentum conserving for unconstrained problems in linear elasticity [1], can exhibit energy explosions and instabilities for contact problems – even within the context of linear elasticity [9]. To resolve this problem, [6] have introduced a contact-implicit variant of the Newmark scheme, with entirely implicit treatment of the contact forces. For the case of linear elasticity, this scheme can be seen to be energy dissipative.

Interestingly, a stable behavior of the energy is not sufficient for excluding spurious numerical results. In fact, even for energy conserving or dissipative schemes oscillations at the contact boundary in velocities and contact forces can occur. For example, although dissipative, the contact-implicit version of the Newmark scheme produces oscillations both in the velocities and in the contact forces during persistent contact, see [3, 6, 7]. In order to tackle this problem, [3] added a projection step to the scheme, whose aim is to provide an admissible predictor, before the force-balancing equation is solved. This removes the boundary oscillations and gives rise to a dissipative scheme, see [3].

However, this stabilized Newmark scheme cannot be directly transferred to the case of contact problems in nonlinear elasticity with nonlinear penetration constraints. Firstly, the proofs for energy dissipativity heavily rely on linearity. Secondly, the Newmark scheme might exhibit instabilities for problems in nonlinear elasticity even in the absence of contact. Therefore, here we consider a time discretization, that combines the ideas for contact stabilization for linear problems from [3, 6, 7] with an approximate "algorithmic" stress field, which has been presented in [9,10] for the case of general hyperelastic materials but without contact constraints.

We formulate this combined scheme for the general case of nonlinear nonpenetration constraints and investigate its behavior numerically. As our numerical results show, our combined scheme is behaving in a stable manner with respect to the contact forces and shows a dissipative behavior of the energy for the investigated examples.

2 Dynamic Contact Problems for Non-linear Material Laws

We are interested in the displacements $\boldsymbol{u} : [t_0, t_{\text{end}}] \times \Omega(t_0) \to \mathbb{R}^3$ of a nonlinear hyperelastic body identified with a bounded (polyhedral) domain $\Omega(t) \subset \mathbb{R}^3$ which might be subjected to volume force densities $\boldsymbol{f} : [t_0, t_{\text{end}}] \times \Omega(t) \to \mathbb{R}^3$, as for instance gravity. We denote Ω as the reference configuration. The nonlinear, hyperelastic, material behavior is characterized by a stored energy function $\boldsymbol{W} : \Omega \times \mathbb{M}^3_+ \to \mathbb{R}$, depending on the space variable and the (right) Cauchy-Green strain tensor $\boldsymbol{C} = \boldsymbol{C}(\boldsymbol{u}) = (\nabla \boldsymbol{u} + \mathrm{id})^T (\nabla \boldsymbol{u} + \mathrm{id})$. If we assume that

$\boldsymbol{W}(\boldsymbol{x},\cdot)$ is continuously differentiable on $\mathbb{M}^3_+$, the first Piola-Kirchhoff stress tensor is then given by $\boldsymbol{T}(\boldsymbol{x},\nabla\boldsymbol{u}) = \frac{\partial}{\partial\nabla\boldsymbol{u}}\boldsymbol{W}(\boldsymbol{x},\nabla\boldsymbol{u})$, cf. [2]. We also define $\hat{\boldsymbol{W}}(\boldsymbol{x},\boldsymbol{C}) := \boldsymbol{W}(\boldsymbol{x},\nabla\boldsymbol{u})$, the stored energy function depending on the (right) Cauchy-Green strain tensor rather than the space derivative of the displacements, and $\hat{\boldsymbol{T}}(\boldsymbol{x},\boldsymbol{C}) = \frac{\partial}{\partial\boldsymbol{C}}\hat{\boldsymbol{W}}(\boldsymbol{x},\boldsymbol{C})$ which is the corresponding Second Piola-Kirchoff tensor [2]. Here, we assume in particular that $\hat{\boldsymbol{W}}$ is a frame-invariant hyperelastic stored energy function.

The boundary $\partial\Omega(t)$ is decomposed into two disjoint parts: the Neumann boundary $\Gamma_N(t)$ where surface tractions $\boldsymbol{p} : [t_0, t_{\text{end}}] \times \Gamma_N(t) \to \mathbb{R}^3$ are given and the potential contact boundary Γ_C, where the body is exposed to non-penetration constraints. Note that for the ease of presentation we do not consider Dirichlet boundaries. Here and in the remainder we denote with $\boldsymbol{n}(\boldsymbol{x})$ the outer normal at $\boldsymbol{x} \in \partial\Omega$.

The deformations are restricted by a rigid obstacle $O \subset \mathbb{R}^3$, which is assumed to be sufficiently smooth. To describe our constraint we employ a signed distance function defined as:

$$d_O(\boldsymbol{u}(\boldsymbol{x})) : \mathbb{R}^3 \to \mathbb{R} \stackrel{\text{def}}{=} \operatorname{dist}(\boldsymbol{u}(\boldsymbol{x}) + \boldsymbol{x}, O^C) - \operatorname{dist}(\boldsymbol{u}(\boldsymbol{x}) + \boldsymbol{x}, O) \tag{1}$$

The non-penetration condition then can be rewritten as

$$d_O(\boldsymbol{u}(\boldsymbol{x})) \le 0 \qquad \text{a.e. on } \Gamma_C \times I \tag{2}$$

[8].

Now, the strong formulation for the dynamic, frictionless contact problem in the reference configuration reads as

$$\ddot{\boldsymbol{u}} - \operatorname{div}\hat{\boldsymbol{T}}(\boldsymbol{C}(\boldsymbol{u})) = \boldsymbol{f} \qquad \text{in } \Omega \times [t_0, t_{\text{end}}] \tag{3a}$$

$$\hat{\boldsymbol{T}}_n(\boldsymbol{C}(\boldsymbol{u})) = \boldsymbol{p} \qquad \text{a.e. on } \Gamma_N \times [t_0, t_{\text{end}}] \tag{3b}$$

$$d_O(\boldsymbol{u}) \le 0 \qquad \text{a.e. on } \Gamma_C \times [t_0, t_{\text{end}}] \tag{3c}$$

$$\hat{\boldsymbol{T}}_n(\boldsymbol{C}(\boldsymbol{u})) \le 0 \qquad \text{a.e. on } \Gamma_C \times [t_0, t_{\text{end}}] \tag{3d}$$

$$\hat{\boldsymbol{T}}_T(\boldsymbol{C}(\boldsymbol{u})) = 0 \qquad \text{a.e. on } \Gamma_C \times [t_0, t_{\text{end}}] \tag{3e}$$

$$d_O(\boldsymbol{u})\hat{\boldsymbol{T}}_n(\boldsymbol{C}(\boldsymbol{u})) = 0 \qquad \text{a.e. on } \Gamma_C \times [t_0, t_{\text{end}}] \tag{3f}$$

$$\boldsymbol{u}(t_0, \boldsymbol{x}) = \boldsymbol{u}^0(\boldsymbol{x}), \quad \dot{\boldsymbol{u}}(t_0, \boldsymbol{x}) = \dot{\boldsymbol{u}}^0(\boldsymbol{x}) \qquad \text{a.e. in } \Omega \tag{3g}$$

Here $\boldsymbol{u}_n = \boldsymbol{u} \cdot \boldsymbol{n}$ denotes the solution in normal direction. $\hat{\boldsymbol{T}}_n = \hat{\boldsymbol{T}}_n(\boldsymbol{u}) = \boldsymbol{n} \cdot \hat{\boldsymbol{T}} \cdot \boldsymbol{n}$ and $\hat{\boldsymbol{T}}_T(\boldsymbol{u}) = \hat{\boldsymbol{T}} \cdot \boldsymbol{n} - \hat{\boldsymbol{T}}_n$ denote the normal and tangential stresses respectively. Furthermore, we define the set of admissible solutions (to be understood in a weak sense):

$$\mathscr{K} = \{\boldsymbol{v} \in H^1(\Omega) | d_O(\boldsymbol{v}) \le 0) \tag{4}$$

Using Green's formula [2], we can formally derive a weak formulation: for every $t \in [t_0, t_{\text{end}}]$, find $\boldsymbol{u}(\cdot, t) \in \mathscr{K}$ s.t.

$$\int_\Omega \ddot{\boldsymbol{u}} \cdot \boldsymbol{v}\, dx + \int_\Omega \hat{\boldsymbol{T}}(\boldsymbol{C}(\boldsymbol{u})) : \nabla \boldsymbol{v} = \int_\Omega \boldsymbol{f} \cdot \boldsymbol{v}\, dx + \int_{\Gamma_N} \boldsymbol{p} \cdot \boldsymbol{v}\, da + \int_{\Gamma_C} \hat{\boldsymbol{T}}_n(\boldsymbol{C}(\boldsymbol{u})) \cdot \boldsymbol{v}\, da, \quad \forall \boldsymbol{v} \in \mathscr{K} \tag{5}$$

In order to simplify our notation, we introduce the external forces:

$$\boldsymbol{F}_{\text{ext}}(\boldsymbol{u}) = \int_\Omega \boldsymbol{f}(t, \boldsymbol{x}) \cdot \boldsymbol{u} + \int_{\Gamma_N} \boldsymbol{p}(t, \boldsymbol{x}) \cdot \boldsymbol{u} \tag{6}$$

as the sum of the volume forces and surface tractions. Similarly, for any displacement $\boldsymbol{u}$, the internal forces are

$$\boldsymbol{F}_{\text{int}}(\boldsymbol{u}) = \int_\Omega \hat{\boldsymbol{T}}(\boldsymbol{C}(\boldsymbol{u})) : \nabla \boldsymbol{u} \tag{7}$$

and the contact forces are defined as:

$$\boldsymbol{F}_{\text{con}}(\boldsymbol{u}) = \int_{\Gamma_C} \hat{\boldsymbol{T}}_n(\boldsymbol{C}(\boldsymbol{u})) \cdot \boldsymbol{u} \tag{8}$$

3 A Time Integration Scheme for Non-linear Elastic Problems

To avoid instabilities in the energy behavior of the Newmark scheme for contact problems in linear-elasticity, Kane et al. [6] have introduced a contact-implicit scheme, which treats the contact forces fully implicitly. While being energy dissipative, this scheme produces oscillations in the contact forces and velocities as shown in, e.g., [7]. Fortunately, these oscillations can easily be removed by adding a discrete L^2-projection to the Newmark scheme, cf. [3]. For the case of contact problems in linear elasticity with linearized non-penetration constraints this stabilized scheme can be shown to be energy dissipative.

Let $\tau > 0$ be the time-step size and let $t_i = t_0 + i\tau$ be the i-th time step. Then, by denoting the approximation at $\boldsymbol{u}(t_i)$ with $\boldsymbol{u}^i$, we can write this scheme as:

Contact-stabilized Newmark scheme

find $\boldsymbol{u}^{i+1}_{\text{pred}}, \boldsymbol{u}^{i+1} \in \mathscr{K}$ and $\dot{\boldsymbol{u}}^{i+1} \in (L^2(\Omega))^3$ such that:

(continued)

(continued)

$$\boldsymbol{u}_{\text{pred}}^{i+1} = \boldsymbol{u}^i + \tau \dot{\boldsymbol{u}}^i - \boldsymbol{G}_{\text{con}}(\boldsymbol{u}_{\text{pred}}^{i+1}) \tag{9a}$$

$$\boldsymbol{u}^{i+1} - \boldsymbol{u}_{\text{pred}}^{i+1} = \frac{\tau^2}{2}\left(\boldsymbol{F}_{\text{int}}^{i+\frac{1}{2}} - (\boldsymbol{F}_{\text{ext}} + \boldsymbol{F}_{\text{con}}^{i+1})\right) \tag{9b}$$

$$\dot{\boldsymbol{u}}^{i+1} - \dot{\boldsymbol{u}}^i = \tau \boldsymbol{F}_{\text{int}}^{i+\frac{1}{2}} - \tau(\boldsymbol{F}_{\text{ext}} + \boldsymbol{F}_{\text{con}}^{i+1}) \tag{9c}$$

Here we used the abbreviation $(\cdot)^{i+\frac{1}{2}} = \frac{1}{2}(\cdot)(\boldsymbol{u}^{i+1}) + \frac{1}{2}(\cdot)(\boldsymbol{u}^i)$ for the forces. In Eq. (9a), the term $\boldsymbol{G}_{\text{con}}(\boldsymbol{u}_{\text{pred}}^{i+1})$ ensures the admissibility of the predictor $\boldsymbol{u}_{\text{pred}}^{i+1}$. The computation of the predictor $\boldsymbol{u}_{\text{pred}}^{i+1}$ itself is realized by means of a discrete L^2 projection of $\boldsymbol{u}^i + \tau \dot{\boldsymbol{u}}^i$ onto the set of feasible solutions.

The energy of the system at time step t_i as the sum of the potential and kinetic energy:

$$\mathscr{E}(\boldsymbol{u}^i, \dot{\boldsymbol{u}}^i) = \int_\Omega \left(-\hat{W}(\boldsymbol{C}^i) + \boldsymbol{F}_{\text{ext}} \cdot \boldsymbol{u}^i\right) d\boldsymbol{x} + \int_\Omega \frac{1}{2}(\dot{\boldsymbol{u}}^i, \dot{\boldsymbol{u}}^i) d\boldsymbol{x} \tag{10}$$

where $\boldsymbol{C}^i = \boldsymbol{C}(\nabla \boldsymbol{u}^i)$ is the right Cauchy Green strain tensor at time step i.

Due to (9a)–(9c), the total change in energy can be computed for the contact-stabilized Newmark scheme as

$$\mathscr{E}(\boldsymbol{u}^{i+1}, \dot{\boldsymbol{u}}^{i+1}) - \mathscr{E}(\boldsymbol{u}^i, \dot{\boldsymbol{u}}^i) =$$

$$\underbrace{\int_\Omega \frac{1}{2}\left(\hat{\boldsymbol{T}}(\boldsymbol{C}^i) + \hat{\boldsymbol{T}}(\boldsymbol{C}^{i+1})\right)(\nabla \boldsymbol{u}^{i+1} - \nabla \boldsymbol{u}^i)\, d\boldsymbol{x} - \int_\Omega \left(\hat{W}(\boldsymbol{C}^{i+1}) - \hat{W}(\boldsymbol{C}^i)\right) d\boldsymbol{x}}_{A}$$

$$\underbrace{- \int_\Omega \boldsymbol{F}_{\text{con}}(\boldsymbol{u}^{i+1})(\boldsymbol{u}^{i+1} - \boldsymbol{u}^i)\, d\boldsymbol{x} - \frac{2}{\tau^2}\int_\Omega \boldsymbol{G}_{\text{con}}(\boldsymbol{u}_{\text{pred}}^{i+1})(\boldsymbol{u}^{i+1} - \boldsymbol{u}_{\text{pred}}^{i+1})\, d\boldsymbol{x}}_{B}$$

In the case of linear elasticity, term (A) vanishes, cf. [3], and the change in energy is due only to the contact terms (B). Moreover, by exploiting the boundary conditions at Γ_C and the linearity of the non-penetration condition, this expression can be seen to be always negative, which shows the dissipativity of the above contact stabilized scheme. However, in the non-linear elastic case, term (A) is nonzero, and the net change in total energy can take arbitrary positive or negative values. Therefore the above scheme is neither energy conserving nor energy dissipative in

the case of non-linear hyperelastic materials. Following [9], we therefore introduce a modification of the above scheme and obtain:

Combined Gonzalez Tensor Scheme

$$\boldsymbol{u}^{i+1}_{\text{pred}} = \boldsymbol{u}^{i} + \tau \dot{\boldsymbol{u}}^{i} - \boldsymbol{G}_{\text{con}}(\boldsymbol{u}^{i+1}_{\text{pred}}) \tag{11a}$$

$$\boldsymbol{u}^{i+1} - \boldsymbol{u}^{i+1}_{\text{pred}} = \frac{\tau^2}{2} \nabla \boldsymbol{u}^{i+\frac{1}{2}} \boldsymbol{S}^{i+\frac{1}{2}}_{\text{algo}} - \boldsymbol{F}_{\text{ext}} - \boldsymbol{F}^{i+1}_{\text{con}} \tag{11b}$$

$$\dot{\boldsymbol{u}}^{i+1} - \dot{\boldsymbol{u}} s^{i} = \tau \left(\nabla \boldsymbol{u}^{i+\frac{1}{2}} \boldsymbol{S}^{i+\frac{1}{2}}_{\text{algo}} - \boldsymbol{F}_{\text{ext}} - \boldsymbol{F}^{i+1}_{\text{con}} \right) \tag{11c}$$

where we replaced the internal forces in (9b) by

$$\boldsymbol{S}^{i+\frac{1}{2}}_{\text{algo}}(\boldsymbol{C}^{i}, \boldsymbol{C}^{i+1}) \overset{\text{def}}{=} D\overline{\boldsymbol{W}}(\boldsymbol{C}^{i}, \boldsymbol{C}^{i+1}) \tag{12}$$

here the specific second Piola Kirchhoff stress tensor is given as

$$D\overline{\boldsymbol{W}}(\boldsymbol{C}^{i}, \boldsymbol{C}^{i+1}) \overset{\text{def}}{=} D\hat{\boldsymbol{W}}(\boldsymbol{C}^{i+\frac{1}{2}}) + \frac{\hat{\boldsymbol{W}}(\boldsymbol{C}^{i+1}) - \hat{\boldsymbol{W}}(\boldsymbol{C}^{i}) - D\hat{\boldsymbol{W}}(\boldsymbol{C}^{i+\frac{1}{2}}) : \Delta^{i+\frac{1}{2}}}{\Delta^{i+\frac{1}{2}} : \Delta^{i+\frac{1}{2}}} \Delta^{i+\frac{1}{2}}$$

and the notation $\Delta^{i+\frac{1}{2}} = \boldsymbol{C}^{i+1} - \boldsymbol{C}^{i}$ and $\boldsymbol{C}^{i+\frac{1}{2}} = \frac{1}{2}(\boldsymbol{C}^{i} + \boldsymbol{C}^{i+1})$ is used.

For this scheme, the change in energy depends solely on the contact condition being met (B), therefore the scheme is energy conserving in absence of contact. With the additional assumption of a linearized constraint, the above scheme is dissipative during contact. We refer to [4] where a detailed analysis is given.

Now we are interested in relaxing this assumption, and testing the scheme on non-linear elasticity problems with non-linear constraints. Even though the scheme is not energy dissipative in this case, the change in energy should be small and we expect the scheme to perform reasonably well.

4 Results

In the following we present numerical results for an application of the combined scheme to unilateral contact problems with non-linear constraints. In the first example we will see how the non-linearity of the constraint affects energy conservation in practical terms. In the second example we will look at the effect of the contact-stabilization step on artificial oscillations examples we used the following stored

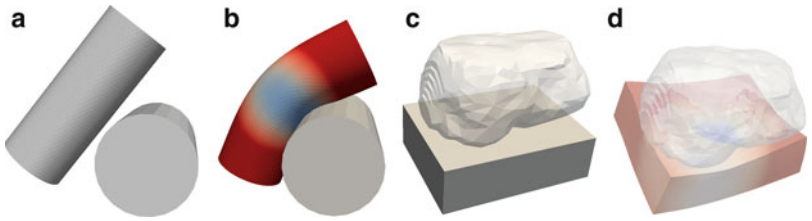


Fig. 1 Displacements depicted for Examples I and II. (**a**) t=0. (**b**) t=120. (**c**) t=0. (**d**) t=250

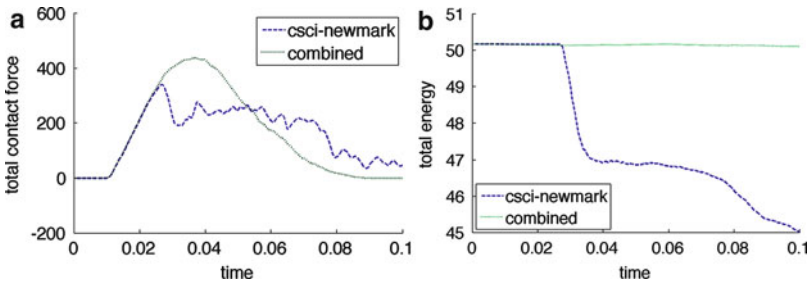


Fig. 2 Comparision of the combined Gonzalez tensor scheme with the contact-stabilized Newmark for Example I. (**a**) Total contact forces. (**b**) Total energy

energy function to describe a non-linear, hyper-elastic, homogenous material model:

$$\hat{\boldsymbol{W}}(\boldsymbol{x}, \boldsymbol{C}) = \int_{\Omega} a \cdot \mathrm{tr}\boldsymbol{E} + b \cdot (\mathrm{tr}\boldsymbol{E})^2 + c \cdot (\mathrm{tr}\boldsymbol{E}^2) + d \cdot \Gamma(\det(I + \nabla \boldsymbol{u})) \quad (13)$$

Here $\boldsymbol{E}(\boldsymbol{u}) = \frac{1}{2}(\boldsymbol{C}(\boldsymbol{u}) - I)$ is the Green-St.Venant strain tensor, $\Gamma(\delta) = c\delta^2 - d \log \delta$ is a logarithmic barrier function and a, b and c are chosen as:

$$a = -d \cdot \Gamma'(1), \quad b = \frac{1}{2}(\lambda - d \cdot -\Gamma'(1) + \Gamma''(1)), \quad c = \nu + d \cdot \Gamma'(1) - \nu \quad (14)$$

where λ and ν are the Lame parameters [2].

In Example I an elastic cylinder moves against a rigid cylindrical polyhedron. The domain of the elastic cylinder is discretized by a tetrahedral mesh. Young's modulus is set to $E = 7{,}000$ while the Poisson ratio to $\nu = 0.2$, and the time step $\tau = 5 \cdot 10^{-4}$. For the simulation we computed 200 time steps. In Fig. 2 we can see a compharision between the combined Gonzalez tensor scheme and the contact-stabilized Newmark scheme. The former performs well: the energy is almost completely conserved, and the evolution of the contact forces looks more natural.

In Example II we chose an application from bio-mechanics: a rigid knee rotula against elastic tissue in the form of a block. The knee rotula is taken from a bone scan. The domain of the elastic tissue is discretized by a hexahedral mesh. The

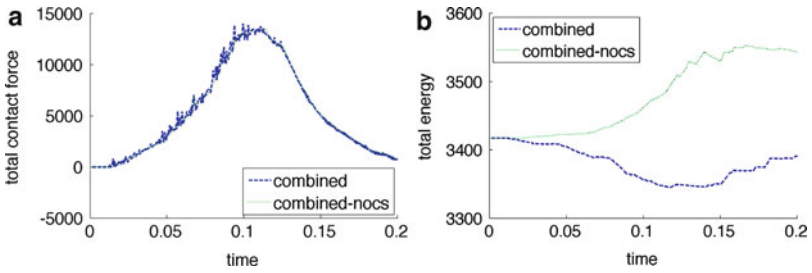


Fig. 3 Comparision of the combined special tensor scheme with and without contact stabilization for Example II. (**a**) Total contact forces. (**b**) Total energy

elasticity parameters are the same as in the previous example, and the time step is $\tau = 5 \cdot 10^{-4}$. The simulation is carried out for 200 time steps. In Fig. 3 we can see a compharision between the combined special tensor scheme with and without the predictor step (11a). Oscillations in the contact forces can be clearly seen in absence of the predictor step. In both cases energy is both dissipated and augmented during the evolution of the problem: however, the net amount is small compared to the initial total energy.

5 Conclusion

From the above numerical results we conclude that the combined, special tensor discretization scheme for contact problems in nonlinear elasticity is not energy dissipative during contact, but is energy conserving in absence of contact as predicted. Moreover the change in energy due to changes at the contact boundary is small relative to the total energy and the behavior of the total energy looks more natural compared to the contact-stabilized Newmark scheme, while also being free of oscillations. Thus the combined scheme shows reasonable behavior for contact problems in non-linear elasticity with non-linear non-penetration constraints.

References

1. Klaus-Juergen Bathe. Conserving energy and momentum in nonlinear dynamics: A simple implicit time integration scheme. 2006.
2. P. G. Ciarlét. Mathematical elasticity, volume I: Three-dimensional elasticity. *Studies in Mathematics and its Applications*, 20(186):715–716, 1988.
3. P. Deuflhard, R. Krause, and S. Ertel. A contact–stabilized newmark method for dynamical contact problems. *International Journal for Numerical Methods in Engineering*, 73(9):1274–1290, 2008. Available as INS Preprint No 0602.
4. Christian Gross, Rolf Krause, and Valentina Poletti. A Contact Stabilized, Energy Dissipative Time Integration Scheme for Non-linear Elasticity Contact Problems. *in preparation*.

5. C. Kane, J.E. Marsden, M. Ortiz, and M. West. Variational integrators and the Newmark algorithm for conservative and dissipative mechanical systems. *International Journal for Numerical Methods in Engineering*, 49:1295–1325, 2002.
6. C. Kane, E. A. Repetto, M. Ortiz, and J. E. Marsden. Finite element analysis of nonsmooth contact. *Computer Methods in Applied Mechanics and Engineering*, 180(1–2):1–26, 1999.
7. Rolf Krause and Mirjam Walloth. Presentation and Comparison of Selected Algorithms for Dynamic Contact Based on the Newmark Scheme. *Applied Numerical Mathematics*, 2009.
8. Stanley J. Osher and Ronald P. Fedkiw. *Level Set Methods and Dynamic Implicit Surfaces*. Springer, 2002.
9. J. C. Simo O. Gonzalez. On the Stability of Symplectic and Energy-Momentum Algorithms for Nonlinear Hamiltonian Systems with Symmetry. *Comput. Methods Appl. Mech. Eng.*, 134:197–222, 1996.
10. Simo O. Gonzalez. Exact energy and momentum conserving algorithms for general models in nonlinear elasticity. *Comput. Methods Appl. Mech. Eng.*, 190(13–14):1763–1783, 2000.

Numerical Simulation of Anisotropic Surface Diffusion of Graphs

D.H. Hoang and M. Beneš

Abstract The aim of this contribution is the numerical simulation of anisotropic surface diffusion of graphs in the context of the epitaxial growth of quantum dots. The numerical scheme is based on the method of lines where the spatial derivatives are approximated by finite differences (Beneš, Appl Math, 48:437–453, 2003). We then solve the resulting ODE system by means of the adaptive Runge-Kutta-Merson method. Finally, we show computational results with various anisotropy settings leading to singular behaviour.

1 Introduction

Quantum dots are nanometre-size semiconductor structures in which the presence or absence of a quantum electron can be used to store information [4]. They were discovered at the beginning of the 1980s and have been a subject of great interest since then. Due to unusual properties, quantum dots are widely used in optical and optoelectronic devices, quantum computation, or biology.

There are many ways to produce quantum dots. Here, we focus on self-assembled quantum dots which appear during molecular beam epitaxy. Due to the effects of mismatch stress, the flat film surface is unstable to small perturbations and may undergo a morphological instability, known as the Asaro-Tiller-Grinfeld instability [7]. The physical mechanism of this instability can be explained as follows. While a flat surface has the lowest surface free energy, a corrugated surface has lower elastic

D.H. Hoang (✉) · M. Beneš
Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University, Trojanova 13, 120 00 Praha 2, Prague, Czech Republic
e-mail: hoangdieu@fjfi.cvut.cz; Michal.Benes@fjfi.cvut.cz

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_19,

energy than the flat one. The elastic energy is lowered by elastic deformation so that the film breaks into the isolated islands (called quantum dots). Therefore, the quantum dots are caused by the competition between surface and elastic energies; elastic energy is reduced as the surface area increases. We assume that the mass is transported by surface diffusion. The surface morphology is also affected by anisotropy in surface energy.

A number of continuum models has been developed for modelling epitaxial growth. A continuum model was derived for the evolution of an epitaxially strained dislocation-free solid film on a rigid substrate by Spencer et al. [6] and on a deformable substrate by Tekalign and Spencer [8]. Xiang and E [9] derived a nonlinear approximation equation for the surface morphology of an infinitely thick stressed solid in 2D.

The work [2] deals with the surface diffusion equation accounting isotropic surface energy (represented by isotropic mean curvature) which leads to smooth surfaces. However, experimental studies of thin films show faceting of surfaces caused by surface energy anisotropy. Hence, the main improvement of this work is the incorporation of the anisotropic mean curvature based on the Finsler geometry into the surface diffusion equation. The numerical studies demonstrate the effect of the surface energy anisotropy on self-assembled growth of quantum dots.

2 Problem Formulation

We consider a vertical two-dimensional cut of the layer where the elastic strain energy density is a given function and vapour deposition is absent.

The film changes its shape because atoms migrate from their original position to another by surface diffusion [5]. The reason is the variation in chemical potential μ which can be viewed as a function over the surface

$$\mu = \Omega(f - \gamma_s \kappa),$$

where Ω is the atomic volume, f is the elastic strain energy density, γ_s is the surface energy, and $\kappa = \kappa_1 + \kappa_2$ is the sum of the local principal curvatures of the surface.

The mass flux $\mathbf{j_s}$ is given by

$$\mathbf{j_s} = -\frac{D_s c_s}{kT} \nabla_s \mu,$$

where D_s is the surface diffusivity, c_s is the concentration of diffusing species, k is the Boltzmann constant, and T is the absolute temperature. The operator ∇_s is the surface gradient operator.

The normal velocity of the surface is

$$v = -\Omega \nabla_s \cdot \mathbf{j_s} = \frac{D_s c_s \Omega^2}{kT} \Delta_s (f - \gamma_s \kappa).$$

where $\Delta_s = \nabla_s \cdot \nabla_s$ is the Laplace-Beltrami operator.

It is convenient to formulate the equation in the dimensionless form. The characteristic length scale l and time scale τ are defined as

$$l = \frac{\gamma_s}{F_0}, \tau = \frac{kT\gamma_s^3}{D_s c_s \Omega^2 F_0^4}.$$

We express the dimensional variables x, y, z, t, f, v, and κ as follows

$$x = Xl, y = Yl, z = Zl, t = T\tau, \kappa = H/l, f = F_0 F,$$

where F_0 is the initial strain energy density. The dimensionless form of the equation is then

$$V = \Delta_s (F - H). \tag{1}$$

3 Graph Case

In what follows we shall study the surface evolution as the graph of the height function. We follow [3] but we assume that the surface energy is anisotropic according to the approach published first in [1].

In order to incorporate the anisotropy into the model we replace the isotropic Euclidean norm in $\mathbb{R}^2$ by another norm exhibiting the desired anisotropy. Following [2] we consider a nonnegative function $\Phi : \mathbb{R}^2 \to \mathbb{R}_0^+$ which is smooth, strict convex, $\mathscr{C}^2(\mathbb{R}^2 \backslash \{0\})$ and satisfying

$$\Phi(t\eta) = |t|\Phi(\eta), \quad t \in \mathbb{R}, \quad \eta \in \mathbb{R}^2,$$
$$\lambda|\eta| \leq \Phi(\eta) \leq \Lambda|\eta|,$$

where $\lambda, \Lambda > 0$. The function given by

$$\Phi^0(\eta^*) = \sup\{\eta^* \cdot \eta | \Phi(\eta) \leq 1\}$$

is its dual. They satisfy the relations

$$\Phi_\eta^0(t\eta^*) = \frac{t}{|t|}\Phi_\eta^0(\eta^*), \qquad \Phi_{\eta\eta}^0(t\eta^*) = \frac{1}{|t|}\Phi_{\eta\eta}^0(\eta^*), \qquad t \in \mathbb{R} - \{0\},$$
$$\Phi(\eta) = \Phi_\eta(\eta) \cdot \eta, \qquad \Phi^0(\eta^*) = \Phi_\eta^0(\eta^*) \cdot \eta^*, \ \eta, \eta^* \in \mathbb{R}^2,$$

where the index η means the derivative with respect to η (i.e., $\Phi^0_\eta = (\partial_{\eta_1}\Phi^0, \partial_{\eta_2}\Phi^0)$). We define the map $T^0 : \mathbb{R}^2 \to \mathbb{R}^2$ as

$$T^0(\eta^*) := \Phi^0(\eta^*)\Phi^0_\eta(\eta^*) \quad \text{for } \eta^* \neq 0,$$
$$T^0(0) := 0.$$

It allows to define the Φ-gradient of a smooth function u as follows

$$\nabla_\Phi u := T^0(\nabla u) = \Phi^0(\nabla u)\Phi^0_\eta(\nabla u) = [T^0_1(\nabla u), T^0_2(\nabla u)],$$

where $\nabla = [\partial_x, \partial_y]$. We assume that there is a function $P : \mathbb{R}^{1+1} \to \mathbb{R}$ such that

$$\Gamma(t) = \{[x, y] \in \mathbb{R}^2 \mid y = P(t, x) \in (a, b)\}.$$

Let $U(x, y) = P(t, x) - y = 0$. Then the anisotropic mean curvature is given by

$$H = \nabla \cdot \left(\frac{\nabla_\Phi U}{\Phi^0(\nabla U)} \right) = \nabla \cdot \left(\frac{T^0(\nabla U)}{\Phi^0(\nabla U)} \right) = \nabla \cdot \left(\frac{T^0(\partial_x P, -1)}{\Phi^0(\partial_x P, -1)} \right)$$
$$= \partial_x \left(\frac{T^0_1(\partial_x P, -1)}{\Phi^0(\partial_x P, -1)} \right).$$

Other quantities are expressed as follows

$$Q(\partial_x P) = \sqrt{1 + |\partial_x P|^2} \qquad \text{(area element)},$$
$$\mathbf{N} = [N, N_2] = \left[\frac{\partial_x P}{Q(\partial_x P)}, \frac{-1}{Q(\partial_x P)} \right] \qquad \text{(normal vector)},$$
$$V = \frac{1}{Q(\partial_x P)} \frac{\partial P}{\partial t} \qquad \text{(normal velocity)}.$$

By substituting these quantities into the Eq. (1) and following [3] we obtain the evolution equations

$$\frac{\partial P}{\partial t} = \partial_x (Q(\partial_x P)(\partial_x (F - H) - (\partial_x (F - H) \cdot N)N)) \text{ on } (a, b) \times (0, T), \quad (2)$$

$$H = \partial_x \left(\frac{T^0_1(\partial_x P, -1)}{\Phi^0(\partial_x P, -1)} \right) \qquad \text{on } (a, b) \times (0, T). \quad (3)$$

The boundary and initial conditions are given by

$$\partial_x P = 0, \quad \partial_x H = 0 \quad \text{on } \{a, b\} \times (0, T),$$
$$P|_{t=0} = P_{ini} \qquad \text{on } [a, b].$$

4 Numerical Scheme

We employed a numerical scheme based on the method of lines. The spatial derivatives are discretized first and the time variable is left continuous. We consider the computational domain (a, b). After discretizing the problem by finite differences in space, we solve the resulting ODE system by the adaptive Runge-Kutta-Merson method. For this purpose, we introduce the following notation:

$$
\begin{aligned}
h &= \frac{b-a}{N} && \text{(mesh size)},\\
u_i &= u(a+ih), &&\\
\omega_h &= \{a+ih | i = 1,\ldots,N-1\} && \text{(grid of internal nodes)},\\
\bar{\omega}_h &= \{a+ih | i = 0,\ldots,N\} && \text{(grid of all nodes)},\\
\gamma_h &= \{a,b\}, &&\\
u_{x,i} &= \frac{u_{i+1}-u_i}{h} && \text{(forward difference)},\\
u_{\bar{x},i} &= \frac{u_i-u_{i-1}}{h} && \text{(backward difference)},\\
\mathscr{P}_h g &= g|_{\bar{\omega}_h} && \text{(projection operator)}.
\end{aligned}
$$

Then we propose a semi-discrete scheme [2]

$$
\begin{aligned}
\frac{\partial P^h}{\partial t} &= (Q(P^h_{\bar{x}})((F-H^h)_{\bar{x}} - ((F-H^h)_{\bar{x}} \cdot N^h)N^h))_x,\\
H^h &= \left(\frac{T_1^0(P^h_{\bar{x}},-1)}{\Phi^0(P^h_{\bar{x}},-1)}\right)_x, \quad N^h = \frac{P^h_{\bar{x}}}{Q(P^h_{\bar{x}})}.
\end{aligned}
$$

The boundary and initial conditions are written as follows

$$
\begin{aligned}
P^h_{\bar{x},0} = P^h_{\bar{x},N} = 0, \quad H^h_{\bar{x},0} = H^h_{\bar{x},N} = 0 \quad &\text{in } (0,T),\\
P^h\Big|_{t=0} = \mathscr{P}_h P_{ini} \quad &\text{on } \bar{\omega}_h.
\end{aligned}
$$

5 Numerical Results

The purpose of this section is to present the numerical results for two types of anisotropy

Table 1 Experimental order of convergence for the $\Phi_1^0(\eta_1, \eta_2)$

N	h	Error L_2	EOC L_2	Error L_∞	EOC L_∞
25	0.08	0.11784	–	0.26666	–
50	0.04	0.04800	1.29580	0.12223	1.12540
100	0.02	0.01008	2.25170	0.03404	1.84430
200	0.01	0.00256	1.97970	0.01068	1.67230

Table 2 Experimental order of convergence for the $\Phi_2^0(\eta_1, \eta_2)$

N	h	Error L_2	EOC L_2	Error L_∞	EOC L_∞
25	0.08	0.05843	–	0.12461	–
50	0.04	0.03173	0.88097	0.09001	0.46926
100	0.02	0.01172	1.43720	0.05212	0.78825
200	0.01	0.00551	1.08850	0.03749	0.47533

$$\Phi_1^0(\eta_1, \eta_2) = \sum_{i=1}^{2} \sqrt{\eta_i^2 + 0.02 \sum_{j=1}^{2} \eta_j^2}$$

and

$$\Phi_2^0(\eta_1, \eta_2) = \sqrt{(\eta_1^2 + \eta_2^2)^2 + 0.02 \sum_{j=1}^{2} \eta_j^2} + \sqrt{(\eta_1^2 - \eta_2^2)^2 + 0.02 \sum_{j=1}^{2} \eta_j^2}.$$

First, we investigate the convergence of the numerical scheme. Then, we explore the long time behaviour of evolution equations (2) and (3) and show the effect of the surface energy anisotropy. In all computations, we consider the initial condition $P_{ini}(x) = 1 + 0.1\cos(2\pi x)$ on the interval $(0, 2)$; the forcing term is set as $F(P) = 100/P$.

Experimental order of convergence. The computations have been performed over a range of different grid resolutions which allows to quantify the numerical convergence by the experimental order of convergence (EOC). Given errors Error_1 and Error_2 for two mesh sizes h_1, h_2, respectively, the EOC is defined as

$$\text{EOC} := \frac{\log(\text{Error}_1/\text{Error}_2)}{\log(h_1/h_2)}.$$

The results are shown in Tables 1 and 2.

Morphology evolution. The solutions at different times for the anisotropy $\Phi_1^0(\eta_1, \eta_2)$ are displayed in Fig. 1. In this example, we observe the development of facets and the surface evolves towards a rectangle-like morphology.

For the anisotropy $\Phi_2^0(\eta_1, \eta_2)$, the obtained evolution is shown in Fig. 2. Faceting occurs again and the surface evolves towards a pyramidal morphology.

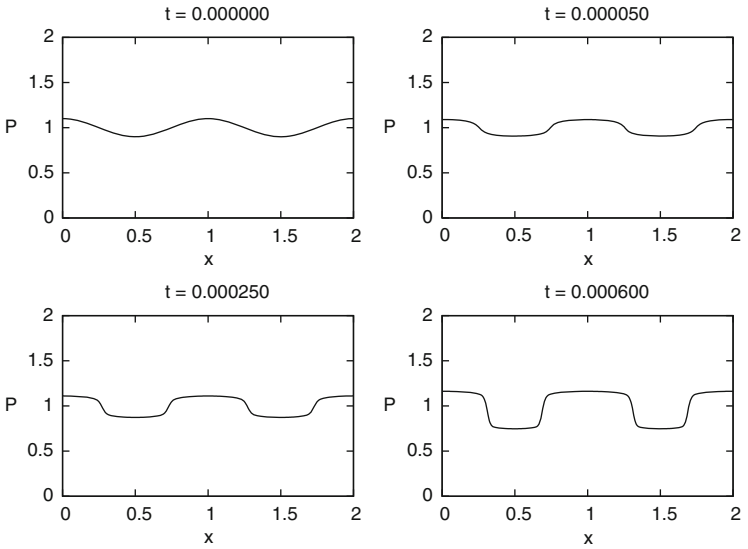


Fig. 1 Rectangle-like morphology

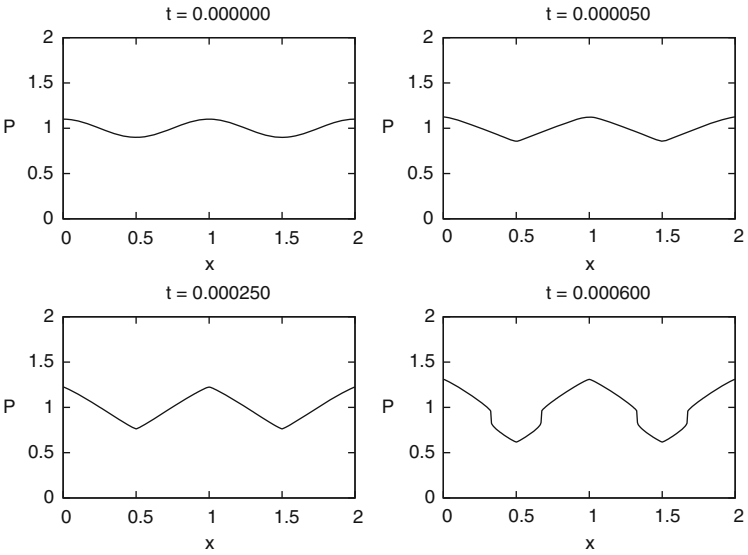


Fig. 2 Pyramidal morphology

6 Conclusion

The simulations showed the influence of surface energy anisotropy on the dynamics of the heteroepitaxial growth leading to the development of faceted rectangle-like and pyramid-like structures. The obtained computational results are similar to the

experimental observation. Next step would be the incorporation of elastic effects generated by the misfit strain between the substrate and the heteroepitaxial film. In order to compute the elastic energy density, the elasticity problem has to be solved or a nonlinear approximation [6, 9] has to be used.

Acknowledgements This work was partially supported by the project "Applied Mathematics in Technical and Physical Sciences" of the Ministry of Education, Youth and Sports of the Czech Republic No. MSM 6840770010 and by the project "Advanced Supercomputing Methods for Implementation of Mathematical Models" of the Student Grant Agency of the Czech Technical University in Prague No. SGS11/161/OHK4/3T/14.

References

1. G. Bellettini and M. Paolini, *Anisotropic motion by mean curvature in the context of finsler geometry*, Hokkaido Mathematical Journal, 25 (1996), pp. 537–566.
2. M. Beneš, *Diffuse-interface treatment of the anisotropic mean-curvature flow*, Applications of Mathematics, 48 (2003), pp. 437–453.
3. ———, *Numerical solution for surface diffusion on graphs*, in Proceedings of the Czech-Japanese Seminar in Applied Mathematics 2005, M. Beneš, M. Kimura, and T. Nakaki, eds., vol. 3 of COE Lecture Note, 2005, pp. 9–25.
4. D. Bimberg, M. Grundmann, and N. Ledentsov, *Quantum dot heterostructures*, John Wiley, 1999.
5. L. B. Freund, *Evolution of waviness on the surface of a strained elastic solid due to stress-driven diffusion*, International Journal of Solids and Structures, 32 (1995), pp. 911–923.
6. B. J. Spencer, S. H. Davis, and P. W. Voorhees, *Morphological instability in epitaxially strained dislocation-free solid films: Nonlinear evolution*, Physical Review B, 47 (1993), pp. 9760–9777.
7. D. J. Srolovitz, *On the stability of surfaces of stressed solids*, Acta Metallurgica, 37 (1989), pp. 621–625.
8. W. T. Tekalign and B. J. Spencer, *Evolution equation for a thin epitaxial film on a deformable substrate*, Journal of Applied Physics, 96 (2004), pp. 5505–5512.
9. Y. Xiang and W. E, *Nonlinear evolution equation for the stress-driven morphological instability*, Journal of Applied Physics, 91 (2002), pp. 9414–9423.

A Special Multiwavelet Basis for Unbounded Product Domains

S. Kestler

Abstract A multiwavelet basis construction for the interval $(0, 1)$ with the special property that the corresponding wavelet discretization of second order constant coefficient differential operators is sparse, is extended to the realline $\mathbb{R}$ and the half-space $\mathbb{R}_+$. The advantage of these new bases is their very convenient usage within adaptive wavelet schemes applied to operator problems on unbounded domains as performance of these schemes is increased while their implementation is facilitated. The construction is explained and underlined by selected numerical experiments.

1 Introduction

Many operator problems in physics or finance are naturally posed on unbounded domains. To cope with the unboundedness of the underlying domain, several numerical methods have been developed in the past. We refer to [1] for a (short) overview. It was shown in [11] that also *adaptive wavelet schemes* as presented in, e.g., [2,3,8,15] can naturally cope with unbounded domains. In this article, we pick up the idea from [11] for the specific setting where, for given univariate domains $\Omega_i \in \{(0,1), \mathbb{R}_+, \mathbb{R}\}$ ($\mathbb{R}_+ := \{x \in \mathbb{R} : x > 0\}$) for $i \in \{1, \dots, n\}$ and an *unbounded* product domain $\square := \Omega_1 \times \cdots \times \Omega_n$, we consider for $f \in H_0^1(\square)'$ the variational problem of finding $u \in H_0^1(\square)$ such that for all $v \in H_0^1(\square)$

$$\langle v, \mathscr{A}[u] \rangle := \sum_{i,j=1}^{n} a_{ij} \int_{\square} \partial_i v \partial_j u + \sum_{i=1}^{n} b_i \int_{\square} v \partial_i u + c_0 \int_{\square} vu = \langle v, f \rangle, \quad (1)$$

S. Kestler (✉)
Institute for Numerical Mathematics, Helmholtzstraße 18, Ulm University, Ulm, Germany
e-mail: sebastian.kestler@uni-ulm.de

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_20,

where a_{ij}, b_i and $c_0 > 0$ are constants such that $\mathscr{A}$ is *boundedly invertible* and $\langle \cdot, \cdot \rangle$ denotes the duality pairing in $H_0^1(\square) \times H_0^1(\square)'$ (see also [5, Eq. (1.1)]). By equipping $H_0^1(\square)$ with a tensor Riesz wavelet basis $\mathbf{D}\Psi := \{\mathbf{D}_{\underline{\lambda}}\psi_{\underline{\lambda}} : \underline{\lambda} \in \nabla\}$, (1) is equivalent to the $\ell_2(\nabla)$-problem: find $\mathbf{u} \in \ell_2(\nabla)$ such that

$$\mathbf{A}\mathbf{u} = \mathbf{f}, \quad \mathbf{f} := \langle \mathbf{D}\Psi, f \rangle \in \ell_2(\nabla), \quad \mathbf{A} := \langle \mathbf{D}\Psi, \mathscr{A}[\mathbf{D}\Psi] \rangle, \tag{2}$$

where $\mathbf{A} : \ell_2(\nabla) \to \ell_2(\nabla)$ is a boundedly invertible operator (see [2]). When $\mathbf{A}$ is symmetric (2) can be solved by an adaptive wavelet scheme which reads idealized for some sufficiently small $\mu \in (0, 1)$ as follows (see [5, Sect. 1.3]):

Set $\Lambda_0 = \emptyset$, $\mathbf{u}_{\Lambda_0} = \mathbf{0}$.
`for` $i = 0, 1, 2, \ldots$
  Find $\Lambda_{i+1} \supset \Lambda_i$ *such that* $\|(\mathbf{A}\mathbf{u}_{\Lambda_i} - \mathbf{f})|_{\Lambda_{i+1}}\|_{\ell_2} \geq \mu \|\mathbf{A}\mathbf{u}_{\Lambda_i} - \mathbf{f}\|_{\ell_2}$ *where, up to some absolute multiple,* $\#(\Lambda_{i+1} \backslash \Lambda_i)$ *is minimal among all such* $\Lambda_{i+1} \subset \nabla$.
  Solve $\mathbf{A}_{\Lambda_{i+1}}\mathbf{u}_{\Lambda_{i+1}} = \mathbf{f}_{\Lambda_{i+1}}$ *where* $\mathbf{A}_{\Lambda_{i+1}} := \mathbf{A}|_{\Lambda_{i+1} \times \Lambda_{i+1}}$ *and* $\mathbf{f}_{\Lambda_{i+1}} := \mathbf{f}|_{\Lambda_{i+1}}$.
`endfor`

Note that if $\mathbf{A}$ is not symmetric, the scheme can be applied to the *normal equations* $\mathbf{A}^\top\mathbf{A}\mathbf{u} = \mathbf{A}^\top\mathbf{f}$ (see [3]). It can be shown that the Galerkin solutions $\mathbf{u}_{\Lambda_i}$ generated by such a scheme converge at an optimal rate $s > 0$: using the Riesz property of $\mathbf{D}\Psi$, we have $\|u - \mathbf{u}_{\Lambda_i}^\top \mathbf{D}\Psi\|_{H_0^1(\square)} \eqsim \|\mathbf{u} - \mathbf{u}_{\Lambda_i}\|_{\ell_2} \lesssim (\#\Lambda_i)^{-s}$ whenever $u = \mathbf{u}^\top\mathbf{D}\Psi$ is included in the *nonlinear* approximation space $\mathscr{A}^s$ (cf. e.g. [13]):

$$\mathbf{u} \in \mathscr{A}^s := \{\mathbf{v} \in \ell_2(\nabla) : |\mathbf{v}|_{\mathscr{A}^s} := \sup_N N^s \|\mathbf{v} - \mathbf{v}_N\|_{\ell_2} < \infty\}, \quad s > 0. \tag{3}$$

Here, we viewed Ψ formally as a column vector and denoted by $\mathbf{v}_N$ a *best N-term approximation* of $\mathbf{v}$ (e.g., the N largest coefficients in modulus of $\mathbf{v}$). Note that we compute a sequence of *finite* domains $(\operatorname{supp}(\mathbf{u}_{\Lambda_i}^\top \mathbf{D}\Psi))_{i=0,1,2,\ldots} \subset \square$ and so, an *adaptive domain truncation* is naturally incorporated in this method (cf. [11]).

Obviously, the above scheme cannot be performed exactly as, e.g., $\mathbf{A}\mathbf{u}_{\Lambda_i}$ is of *infinite* support since $\mathbf{A}$ is in general *not* sparse and has infinitely many entries per column. To address this issue, one has to show that $\mathbf{A}\mathbf{u}_{\Lambda_i}$ permits w.r.t. a certain target accuracy *finitely* supported *approximations*. This leads to the concept of $s^\star$-*compressibility* (see e.g. [12]). This concept permits the set up of so-called **APPLY** routines (see e.g. [2, 15]) which approximate infinitely supported matrix vector products. For the approximation of $\mathbf{f}$, we refer to [8, 11].

To overcome the usage of an **APPLY**-routine (which is quantitatively demanding and difficult to implement), special multiwavelet bases Ψ have been constructed in [5, 7] which actually lead to $\mathbf{A}$ which is *sparse* when $\mathscr{A}$ has the specific form from (1) and $\square = (0, 1)^n$. Now, $\mathbf{A}\mathbf{u}_{\Lambda_i}$ is *finitely supported* and can be computed *exactly* within linear complexity resulting in a considerable speed-up of the scheme.

We show that these interval bases permit intuitive extensions to $\mathbb{R}$ and $\mathbb{R}_+$. Exemplarily, we consider the construction from [7] and use our new wavelet construction to solve operator problems of type (1) on *unbounded* product domains.

2 Tensor Multiwavelet Basis Construction

Our basis construction for $H_0^1(\square)$ is based on the principle of tensorizing special univariate Riesz (wavelet) bases whose construction will be explained below.

2.1 Tensorization of Univariate Bases

We exploit the fact that (cf. e.g. [9]) for $\square = \Omega_1 \times \cdots \times \Omega_n$,

$$H_0^1(\square) \eqsim \bigcap_{k=1}^{n} \bigotimes_{\ell=1}^{n} H_0^{0+\delta_{k,\ell}}(\Omega_\ell) \quad \text{with} \quad H_0^0(\Omega_i) := L_2(\Omega_i),$$

where $\Omega_i \in \{(0,1), \mathbb{R}_+, \mathbb{R}\}$ for $i \in \{1, \ldots, n\}$ and $H_0^1(0,1) := \operatorname{clos}_{H^1(0,1)} \mathscr{D}(0,1)$, $H_0^1(\mathbb{R}_+) := \operatorname{clos}_{H^1(\mathbb{R}_+)} \mathscr{D}(\mathbb{R}_+)$, $H_0^1(\mathbb{R}) := \operatorname{clos}_{H^1(\mathbb{R})} \mathscr{D}(\mathbb{R}) = H^1(\mathbb{R})$.

The most important step in our tensor basis construction is the set up of suitable univariate multiwavelet bases Υ^Ω for $L_2(\Omega)$ for $\Omega \in \{(0,1), \mathbb{R}_+, \mathbb{R}\}$,

$$\Upsilon^\Omega := \{\psi_\lambda : \lambda := (i,j,k) \in \mathscr{J}^\Omega\}. \tag{4}$$

Here, $\mathscr{J}^\Omega$ is a countable index set, adopted to the underlying domain Ω. We denote by i the *type*, by $|\lambda| := j$ the *level* and by k the *translation index* of the wavelet ψ_λ. Defining $\mathscr{H}^s(\Omega) := [L_2(\Omega), H^4(\Omega) \cap H_0^1(\Omega)]_{s/4}$ for $s \in [0,4]$ and $\mathscr{H}^s(\Omega) := (H^{-s}(\Omega))'$ for $s < 0$, we require (compare [7, Sect. 2]):

1. The properly scaled collection $\{\psi_\lambda / \|\psi_\lambda\|_{\mathscr{H}^s(\Omega)} : \lambda \in \mathscr{J}^\Omega\}$ is a Riesz basis for $\mathscr{H}^s(\Omega)$ for $-\widetilde{\gamma} < s < \gamma$ where $\gamma > 1, \widetilde{\gamma} > 0$ are smoothness indices.
2. Polynomial order: ψ_λ is a piecewise polynomial of order d.
3. Local supports: $\operatorname{diam}(\operatorname{supp} \psi_\lambda) \sim 2^{-|\lambda|}$.
4. Vanishing moments: $\int_{\operatorname{supp} \psi_\lambda} \psi_\lambda \, p = 0$ for all $p \in \mathscr{P}_{d-1}$ when $|\lambda| > 0$.
5. $\int \partial^\alpha \psi_\lambda \partial^\beta \psi_{\lambda'} = (-1)^\alpha \int \psi_\lambda \, \partial^{\alpha+\beta} \psi_{\lambda'} = 0$ for $\alpha, \beta \in \{0,1\}$ and $\big||\lambda| - |\lambda'|\big| > 1$.

Note that property 5 distinguishes Υ^Ω from other possible wavelet basis constructions for Ω (see [15] for an overview) and is crucial to obtain sparsity in $\mathbf{A}$. With Υ^{Ω_i} from (4) satisfying 1–5 for $i \in \{1, \ldots, n\}$, we define the *tensor* basis for $L_2(\square)$:

$$\Psi = \{\psi_{\underline{\lambda}} := \psi_{\lambda_1} \otimes \cdots \otimes \psi_{\lambda_n} : \underline{\lambda} := (\lambda_1, \ldots, \lambda_n) \in \nabla := \mathscr{J}^{\Omega_1} \times \cdots \times \mathscr{J}^{\Omega_n}\}. \tag{5}$$

Renormalizing Ψ in $H_0^1(\square)$, i.e., considering $\mathbf{D}\Psi := \{\mathbf{D}_{\underline{\lambda}} \psi_{\underline{\lambda}} : \underline{\lambda} \in \nabla\}$ with $\mathbf{D} := \operatorname{diag}[\mathbf{D}_{\underline{\lambda}} : \underline{\lambda} \in \nabla]$ and $\mathbf{D}_{\underline{\lambda}} := \|\psi_{\underline{\lambda}}\|^{-1}_{H_0^1(\square)}$ yields a Riesz basis for $H_0^1(\square)$ (cf. [5]). Since $\Psi = \Upsilon^{\Omega_1} \otimes \cdots \otimes \Upsilon^{\Omega_n}$, using properties 3 and 5, $\mathbf{A} = (\mathbf{a}_{\underline{\lambda},\underline{\lambda}'})_{\underline{\lambda},\underline{\lambda}' \in \nabla}$ is sparse since $\mathbf{a}_{\underline{\lambda},\underline{\lambda}'} = 0$ when $\max\{\big||\lambda_1| - |\lambda_1'|\big|, \ldots, \big||\lambda_n| - |\lambda_n'|\big|\} > 1$.

2.2 Construction of the Special Univariate Multiwavelet Bases

For the construction of Υ^Ω satisfying 1–5, we have to define suitable primal and dual *multiresolution spaces* for $j \in \mathbb{N}_0$ (compare [7, Eqs. (6) and (7)])

$$V_j^\Omega := \prod_{k\in\mathbb{Z}} \mathscr{P}_3((k2^{-(j+1)}, (k+1)2^{-(j+1)}) \cap \Omega) \cap C^1(\Omega) \cap H_0^1(\Omega),$$

$$\widetilde{V}_j^\Omega := \prod_{k\in\mathbb{Z}} \mathscr{P}_3((k2^{-j}, (k+1)2^{-j}) \cap \Omega) \cap L_2(\Omega),$$

and to equip them with uniform Riesz bases Φ_j^Ω and $\widetilde{\Phi}_j^\Omega$. By $\mathscr{P}_m(a,b)$, we denote the polynomials on (a,b) up to degree m. Next, one equips the *detail spaces*

$$W_j^\Omega := V_j^\Omega \cap (\widetilde{V}_{j-1}^\Omega)^{\perp_{L_2(\Omega)}} \text{ with Riesz bases } \Upsilon_j^\Omega := \{\psi_\lambda : \lambda \in \mathscr{J}^\Omega, |\lambda| = j\}$$

which are uniformly $L_2(\Omega)$ stable. We then intend to define $\Upsilon^\Omega := \Phi_0 \cup \cup_{j\in\mathbb{N}} \Upsilon_j^\Omega$.

Note that by the specific definition of V_j^Ω and $\widetilde{V}_j^\Omega$, property 5 is satisfied. Indeed, we obviously have $\partial^{\alpha+\beta}\psi_{\lambda'} \subset \widetilde{V}_{|\lambda'|+1}^\Omega$ for $\alpha, \beta \in \{0,1\}$ and for all $\lambda' \in \mathscr{J}^\Omega$. The assertion then follows directly from $\psi_\lambda \perp \widetilde{V}_{|\lambda|-1}^\Omega$ for all $\lambda \in \mathscr{J}^\Omega$ with $|\lambda| > 0$.

For the construction of the wavelets $\psi_\lambda \in W_j^\Omega$, we consider the mother *scaling functions* $\varphi^{(i)}$ ($i \in \{1,2\}$) and the mother wavelets $\psi^{(i)}$ ($i \in \{1,\ldots,4\}$) as defined in [7, Sect. 5] (see also Fig. 1) and define in dependence of Ω:

$$\begin{aligned}\Phi_j^{(0,1)} &:= \{\varphi_{j,0}^{(2)}|_{(0,1)}\} \cup \{\varphi_{j,k}^{(i)} : i \in \{1,2\}, k \in \{1,2,\ldots,2^{j+1}-1\}\} \cup \{\varphi_{j,2^{j+1}}^{(2)}|_{(0,1)}\} \\ &:= \Phi_j^{L,(0,1)} \cup \Phi_j^{I,(0,1)} \cup \Phi_j^{R,(0,1)}, \qquad (6)\\ \Upsilon_j^{(0,1)} &:= \{\psi_{j,0}^{(i)}|_{(0,1)} : i \in \{1,2,4\}\} \cup \{\psi_{j,k}^{(i)} : i \in \{1,\ldots,4\}, k \in \{2,4,\ldots,2^j-2\}\} \\ &\quad \cup \{\psi_{j,2^j}^{(4)}|_{(0,1)}\} := \Upsilon_j^{L,(0,1)} \cup \Upsilon_j^{I,(0,1)} \cup \Upsilon_j^{R,(0,1)}, \qquad (7)\\ \Phi_j^{\mathbb{R}+} &:= \Phi_j^{L,(0,1)} \cup \Phi_j^{I,\mathbb{R}+}, \quad \Phi_j^{I,\mathbb{R}+} := \{\varphi_{j,k}^{(i)} : i \in \{1,2\}, k = 1,2,\ldots\}, \qquad (8)\\ \Upsilon_j^{\mathbb{R}+} &:= \Upsilon_j^{L,(0,1)} \cup \Upsilon_j^{I,\mathbb{R}+}, \quad \Upsilon_j^{I,\mathbb{R}+} := \{\psi_{j,k}^{(i)} : i \in \{1,\ldots,4\}, k = 2,4,\ldots\}, \qquad (9)\\ \Phi_j^{\mathbb{R}} &:= \{\varphi_{j,k}^{(i)} : i \in \{1,2\}, k \in \mathbb{Z}\}, \quad \Upsilon_j^{\mathbb{R}} := \{\psi_{j,k}^{(i)} : i \in \{1,\ldots,4\}, k \in 2\mathbb{Z}\}, \qquad (10)\end{aligned}$$

where $\varphi_{j,k}^{(i)} := 2^{\frac{1}{2}(j+1)}\varphi^{(i)}(2^{j+1}\cdot - k)$ and $\psi_{j,k}^{(i)} := 2^{\frac{1}{2}j}\psi^{(i)}(2^j \cdot - k)$.

The construction principle is simple: concerning $\mathbb{R}_+$, we take the left boundary parts (superscript "L") from the interval construction and extend the inner part (superscript "I") by adding translation indices up to $+\infty$. In particular, there is no right boundary part (superscript "R"). For $\mathbb{R}$, we do not need the left boundary part. We consider all translations of the mother scaling functions $\varphi^{(i)}$ ($i \in \{1,2\}$) and mother wavelets $\psi^{(i)}$ ($i \in \{1,\ldots,4\}$ (translation invariant case).

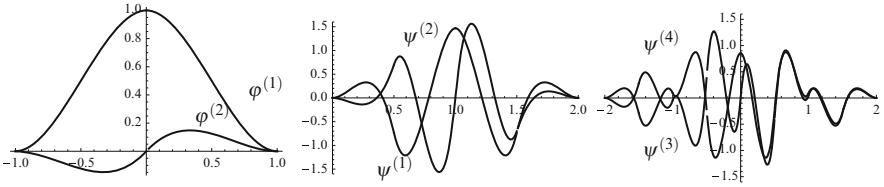


Fig. 1 The (mother) scaling functions $\varphi^{(1)}$, $\varphi^{(2)}$ and the (mother) wavelets $\psi^{(1)}$, $\psi^{(2)}$, $\psi^{(3)}$, $\psi^{(4)}$

Theorem 1. *Let $\Omega \in \{(0,1), \mathbb{R}_+, \mathbb{R}\}$. With Φ_0^Ω and Υ_j^Ω as defined in (6)–(10),*

$$\Upsilon^\Omega := \{\psi_\lambda : \lambda \in \mathscr{J}^\Omega\} = \Phi_0^\Omega \cup \bigcup_{j\in\mathbb{N}} \Upsilon_j^\Omega, \quad \Upsilon^{s,\Omega} := \{\psi_\lambda / \|\psi_\lambda\|_{\mathscr{H}^s(\Omega)} : \lambda \in \mathscr{J}^\Omega\},$$

are stable Riesz (wavelet) bases for $L_2(\Omega)$ and $\mathscr{H}^s(\Omega)$ for $s \in (-\frac{1}{2}, \frac{5}{2})$, respectively. In particular, the elements of Υ^Ω satisfy 1–5.

A full proof of these results would extend the scope of this paper, therefore we only give a sketch of the major relevant parts.

Proof. For $\Omega = (0,1)$, we refer to [7]. For $\Omega \in \{\mathbb{R}_+, \mathbb{R}\}$, the proof is also based on [6, Theorem 2.1]. It is not difficult to prove the necessary *direct* and *inverse* estimates for V_j^Ω, $\widetilde{V}_j^\Omega$ (see [6, Eqs. (C2) and (C3)]). Moreover,

$$\inf_{0\neq v_j\in V_j^\Omega} \sup_{0\neq \widetilde{v}_j\in\widetilde{V}_j^\Omega} \frac{\langle v_j, \widetilde{v}_j\rangle_{L_2}}{\|v_j\|_{L_2}\|\widetilde{v}_j\|_{L_2}} \gtrsim 1, \qquad \inf_{0\neq \widetilde{v}_j\in\widetilde{V}_j^\Omega} \sup_{0\neq v_j\in V_j^\Omega} \frac{\langle \widetilde{v}_j, v_j\rangle_{L_2}}{\|\widetilde{v}_j\|_{L_2}\|v_j\|_{L_2}} \gtrsim 1,$$

with constants independent of j can also be proven for $\Omega \in \{\mathbb{R}_+, \mathbb{R}\}$ using results from [7, Sect. 7]. It remains to show that the collections Υ_j^Ω are uniform Riesz bases for W_j^Ω. To this end, one has to show (uniform) $L_2(\Omega)$ stability of $\Upsilon_j^\Omega = \{\psi_\lambda : \lambda \in \mathscr{J}_j^\Omega\}$ as well as that $\operatorname{clos}_{L_2(\Omega)}\operatorname{span}\Upsilon_j^\Omega = W_j^\Omega$. The last point is delicate. For $\Omega = (0,1)$, it is sufficient to show that $\operatorname{span}\Upsilon_j^\Omega \subset W_j^\Omega$ and that $\dim\Upsilon_j^\Omega = \dim W_j^\Omega = 2^{j+1}$. However, for $\Omega \in \{\mathbb{R}_+, \mathbb{R}\}$, Υ_j^Ω as well as W_j^Ω are both *countable*, but not finite and so, a dimension based argument cannot be applied. To this end, it will be shown in [10] that the assumption that there actually exist $w_j^\Omega \in W_j^\Omega$ and $\varepsilon > 0$ such that $\|w_j^\Omega - \mathbf{d}^\top\Upsilon_j^\Omega\|_{L_2(\Omega)} > \varepsilon$ for all $\mathbf{d} \in \ell_2(\mathscr{J}_j^\Omega)$, i.e., w_j^Ω is not in the $L_2(\Omega)$ closure of the linear span of Υ_j^Ω, leads to a contradiction. □

So far, in the definition of the wavelet bases Υ^Ω, we have fixed the so-called *minimal level* j_0 (the level on which we also have scaling functions $\varphi^{(i)}$ $(i \in \{1,2\})$) at $j_0 = 0$. That it was described in [11] for *unbounded* domains, it is convenient to set up wavelet bases with minimal levels lower than $j_0 = 0$. The following result can now be proven analogously to Theorem 1:

Table 1 Bounds for the Riesz constants from (11). For $L_2(\Omega)$, they do not depend on j_0

	$\Omega = (0,1)$	$\Omega = \mathbb{R}_+$			$\Omega = \mathbb{R}$		
j_0	0	0	-2	-4	0	-2	-4
$c_{L_2(\Omega)}$	0.04	0.04	–	–	0.04	–	–
$C_{L_2(\Omega)}$	2.46	2.46	–	–	2.46	–	–
$c_{H_0^1(\Omega)}$	0.27	0.27	0.25	0.11	0.27	0.25	0.11
$C_{H_0^1(\Omega)}$	2.00	2.00	2.00	2.00	2.00	2.00	2.00

Corollary 1. *Let $\Omega \in \{\mathbb{R}_+, \mathbb{R}\}$ and $T_{j_0}[v] := 2^{\frac{1}{2}j_0} v(2^{j_0}\cdot)$ for $v \in L_2(\Omega)$. Then,*

$$\Upsilon^{(s,j_0),\Omega} := T_{j_0}[\Upsilon^{s,\Omega}] := \{T_{j_0}[\psi_\lambda]/\|T_{j_0}[\psi_\lambda]\|_{\mathscr{H}^s(\Omega)} : \lambda \in \mathscr{J}^\Omega\}$$

is a stable Riesz basis for $\mathscr{H}^s(\Omega)$ for $j_0 \in \mathbb{Z}$ and for $s \in (-\frac{1}{2}, \frac{5}{2})$ satisfying 1–5.

In order to use the constructed wavelet bases $\Upsilon^{(s,j_0),\Omega}$ for $\Omega \in \{(0,1), \mathbb{R}_+, \mathbb{R}\}$ ($\Upsilon^{(s,j_0),(0,1)} \equiv \Upsilon^{s,(0,1)}$) within an adaptive wavelet scheme, bounds for their *Riesz constants* are required. With $V \in \{L_2(\Omega), H_0^1(\Omega)\}$, one has to estimate

$$c_V := \inf_{0\neq \mathbf{d}\in \ell_2(\mathscr{J}^\Omega)} \frac{\|\sum_\lambda d_\lambda \psi_\lambda\|_V^2}{\sum_\lambda |d_\lambda|^2 \|\psi_\lambda\|_V^2}, \quad C_V := \sup_{0\neq \mathbf{d}\in \ell_2(\mathscr{J}^\Omega)} \frac{\|\sum_\lambda d_\lambda \psi_\lambda\|_V^2}{\sum_\lambda |d_\lambda|^2 \|\psi_\lambda\|_V^2}, \quad (11)$$

where $\sum_\lambda := \sum_{\lambda\in\mathscr{J}^\Omega}$ (see [5, Sect. 1.1]). In Table 1, we give bounds for these constants for different values of j_0. Here, we proceeded as in [11, Sect. 3] and set $\|v\|_{H_0^1(0,1)} := \|\partial v\|_{L_2(0,1)}^2$, $\|v\|_{H_0^1(\Omega)}^2 := \|v\|_{L_2(\Omega)}^2 + \|\partial v\|_{L_2(\Omega)}^2$ for $\Omega \in \{\mathbb{R}_+, \mathbb{R}\}$.

Remark 1. By choosing possibly different minimal levels $j_0^{(i)}$ for the different coordinate directions $i \in \{1, \dots, n\}$, we may redefine Ψ from (5) by

$$\Psi_{\mathbf{j}_0} := \Upsilon^{(0,j_0^{(1)}),\Omega_1} \otimes \cdots \otimes \Upsilon^{(0,j_0^{(n)}),\Omega_n}, \quad \mathbf{j}_0 := (j_0^{(1)}, \dots, j_0^{(n)}).$$

The $H_0^1(\square)$ normalized collection $\mathbf{D}\Psi_{\mathbf{j}_0}$ is again a Riesz basis for $H_0^1(\square)$.

Remark 2. In [4], *piecewise* tensor product wavelet bases have been constructed. Future research is needed to investigate the potential of this approach for unbounded product domains in order to treat more complicated unbounded domains.

3 Selected Numerical Results

We consider (1) for $\square_1 := \mathbb{R}^2$ and $\square_2 := (0,1) \times \mathbb{R}_+$ where we set $a_{11} = a_{22} = c_0 = 1$ and all other coefficients to 0. In view of adaptive domain truncation, the first domain type is most challenging. For $\mathbb{R}^2$, singularities in the solution can only arise from the right-hand side whereas for $\square_2$, they can also arise from the domain. We consider continuous reference solutions to (1), $u_1 \in H_0^1(\square_1)$ and $u_2 \in H_0^1(\square_2)$

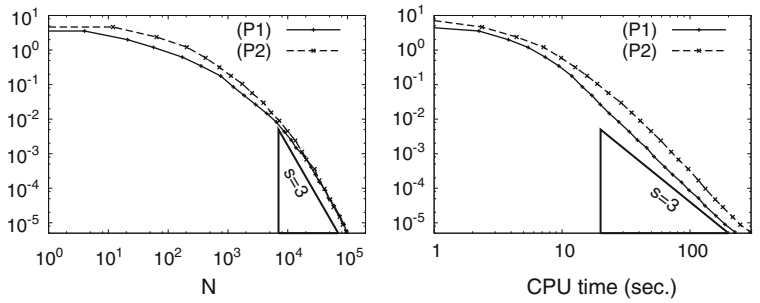


Fig. 2 Approximation error of **AWGM** measured in $\|\cdot\|_{H_0^1(\square)}$ for $\mathbf{j}_0 = (-2,-1)$ for (P1) and $\mathbf{j}_0 = (0,-1)$ for (P2) w.r.t. $N := \#\operatorname{supp} \mathbf{u}_{\Lambda_i}$ (*left*) and w.r.t. the required CPU time (*right*)

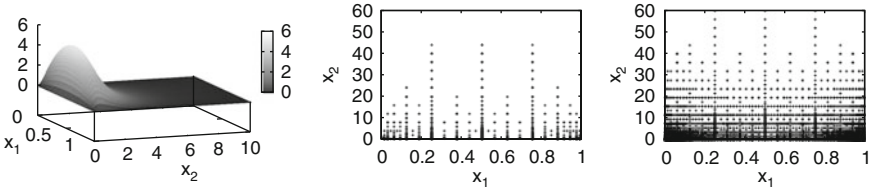


Fig. 3 Solution for (P3) (*left*) and computational domains for $\#\operatorname{supp} \mathbf{u}_{\Lambda_i} = 5,202$ (*middle*), $\#\operatorname{supp} \mathbf{u}_{\Lambda_i} = 51,242$ (*right*). A cross refers to the center of supp $\psi_{\underline{\lambda}}$ for $\underline{\lambda} \in \Lambda_i$

which are singular along $x_1 = \frac{1}{3}$ and $x_2 = \frac{1}{3}$. We also consider f_3 which does vanish not at the corners of $\square_2$:

$$\begin{aligned}
&\text{(P1)}\quad u_1(x_1,x_2) := 0.1\cdot e^{\left(-\frac{1}{10}|x_1-\frac{1}{3}|\right)}\cdot 0.5\cdot e^{\left(-\frac{1}{2}|x_2-\frac{1}{3}|\right)},\\
&\text{(P2)}\quad u_2(x_1,x_2) := 10\cdot(e^{x_1}-1)\cdot \mathbf{1}_{\{0\le x_1<\frac{1}{3}\}}\cdot 10\cdot(e^{-\frac{1}{2}(x_1-1)}-1)\mathbf{1}_{\{\frac{1}{3}\le x_1\le 1\}}\\
&\qquad\qquad \cdot(e^{x_2}-1)\cdot \mathbf{1}_{\{0\le x_2<\frac{1}{3}\}}\cdot e^{-\frac{1}{10}(x_2-\frac{1}{3})}\cdot(e^{\frac{1}{3}}-1)\cdot \mathbf{1}_{\{\frac{1}{3}\le x_2\}},\\
&\text{(P3)}\quad f_3(x_1,x_2) := 10\cdot e^{-x_2}.
\end{aligned}$$

We use the adaptive wavelet (Galerkin) method **AWGM** from [5] with optimized parameters $\mu = 0.6$, $\gamma = 0.15$ that are outside the range for guaranteed convergence. The approximation of $\mathbf{f}$ and $\mathbf{j}_0$ are obtained as in [11, Sect. 5]. For the implementation, the C++ library LAWA [14] is used.

With the univariate wavelet bases being of polynomial order $d = 4$, the best possible rate s (see (3)) for error reduction over the *whole unbounded domain* is bounded by $s \le d-1$. For $s < d-1$, this is shown in [13] for u belonging to the intersection of certain tensor Besov spaces. However, $s = d-1$ is the rate that can be proven for sufficiently smooth functions.

Despite the low Sobolev regularity of the u_i, the best rate is attained within linear complexity (see Fig. 2). The maximum number of CG iterations to approximate $\mathbf{A}\mathbf{u}_{\Lambda_i} = \mathbf{f}_{\Lambda_i}$ was 12. With decreasing target tolerance in the **AWGM**, the computational domain $\operatorname{supp}(\mathbf{u}_{\Lambda_i}^\top \mathbf{D}\Psi_{\mathbf{j}_0})$ is adaptively enlarged and the singularities at the corners of $\square_2$ are resolved (see Fig. 3).

Acknowledgements The author is grateful to the DFG Research Training Group 1100 for financial support and would like to thank his PhD advisor Prof. Dr. Karsten Urban.

References

1. T. Z. Boulmezaoud. Inverted finite elements: A new method for solving elliptic problems in unbounded domains. ESAIM: M2AN, **39**(1), 109–145 (2005).
2. A. Cohen, W. Dahmen, and R. DeVore. Adaptive wavelet methods for elliptic operator equations: convergence rates. Math. Comp., **70**(233), 27–75 (2001).
3. A. Cohen, W. Dahmen, and R. DeVore. Adaptive wavelet methods II – beyond the elliptic case. Found. Comput. Math., **2**, 203–245 (2002).
4. N. Chegini, S. Dahlke, U. Friedrich and R. Stevenson. Piecewise tensor product wavelet bases by extensions and approximation rates. Preprint (2012).
5. N. Chegini and R. Stevenson. The adaptive tensor product wavelet scheme: Sparse matrices and the application to singularly perturbed problems. IMA J. Numer. Anal., **32**(1), 75–104 (2012).
6. W. Dahmen and R. Stevenson. Element-by-element construction of wavelets satisfying stability and moment conditions. SIAM J. Numer. Anal., **37**(1), 319–352 (1999).
7. T. Dijkema and R. Stevenson. A sparse Laplacian in tensor product wavelet coordinates. Numer. Math., **115**, 433–449 (2010).
8. T. Gantumur, H. Harbrecht, and R. Stevenson. An optimal adaptive wavelet method without coarsening of the iterands. Math. Comp., **76**(258), 615–629 (2007).
9. M. Griebel and P. Oswald. Tensor product type subspace splitting and multilevel iterative methods for anisotropic problems. Adv. Comput. Math., **4**, 171–206 (1995).
10. S. Kestler. Adaptive wavelet methods for multi-dimensional problems in numerical finance. PhD thesis (in preparation).
11. S. Kestler and K. Urban. Adaptive wavelet methods on unbounded domains. J. Sci. Comp., **53**(2), 342–376 (2012).
12. C. Schwab and R. Stevenson. Adaptive wavelet algorithms for elliptic PDE's on product domains. Math. Comp., **77**(261), 71–92 (2008).
13. W. Sickel and T. Ullrich. Tensor products of Sobolev-Besov spaces and applications to approximation from the hyperbolic cross. J. Approx. Theory, **161**, 748–786 (2009).
14. A. Stippler. LAWA – Library for Adaptive Wavelet Applications. http://lawa.sourceforge.net (2009).
15. K. Urban. *Wavelet methods for elliptic partial differential equations*. Oxford University Press (2009).

Parameter Estimation Problems in Physically Based Image Processing

M. Klinger

Abstract In this contribution we consider an optimization problem constrained by a system of state equations coupling the nonstationary model for gray-value transport in an image sequence to the physical model of a transport field resulting in the gray-value evolution. Since in this situation the movement over the boundaries is often unknown, we use a Dirichlet-boundary control formulation for the determination of the transport field.

1 Motivation

We want to introduce a real world application to motivate the investigation of optimization problems arising from physically based image processing. Because of the structure of the data and the high complexity of the real world physical model, we will throughout this paper consider a mathematically manageable test situation.

In environmental sciences it is of high interest to determine the movement of substances by the atmospheric wind system. Observations of areas like the polar region or huge deserts suffer from the problem that the installation of a dense grid of measurement systems on the ground is at the moment technically impossible. As a consequence it is desirable to use the data given as image sequences from satellite remote sensing. To fix ideas, we concentrate on the movement of airbourne dust in the earth's atmosphere. The image data is hereby supplied by a geosynchronous

M. Klinger (✉)
Institut für Angewandte Mathematik, Im Neuenheimer Feld 294, 69120, Heidelberg, Germany
e-mail: matthias.klinger@iwr.uni-heidelberg.de

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_21,

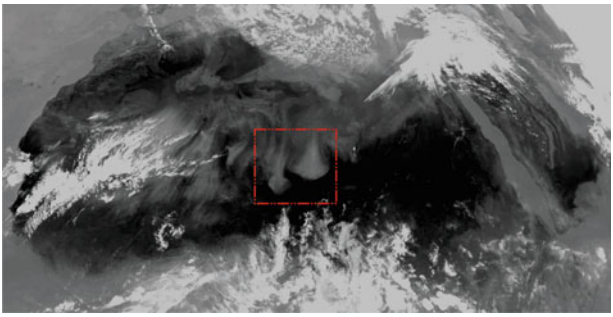


Fig. 1 Gray-value image of the Sahara desert taken by geosynchronous satellite. *Dashed box*: Aperture of a dust event in the original image

satellite (see Fig. 1). Scientists are interested in the sources, the trajectories and of course the deposition of the sand, since sand acts as a nutrient in several local ecosystems, e.g. Amazonas (Schepanski, Tegen and Macke [10]).

The gray-value picture in Fig. 1 represents an observation of the whole Northern Africa. In fact, we only observe dust movement in a small subdomain of this whole image domain (see the dashed box in Fig. 1). For the reduction of computational effort, we therefore truncate the image to interesting apertures of the so called dust events. Since there is gray-value movement over the boundary of the truncated image, we have in general no information on the transport fields on the boundary.

The aim of this article is therefore to present a novel methodology, which considers the following three aspects of the above mentioned real world application:

1. Estimation of **optical flow** for image sequences by **variational** methods,
2. Based on the underlying **physical model** of the observed phenomenon, and
3. Using a **Dirichlet-boundary control** formulation in order to deal with the lack of information of transport fields on the boundary of the computational domain.

2 Image Processing Issues for a Model Situation

The starting point of the following considerations is the transport equation

$$\begin{aligned} \partial_t I(\mathbf{x},t) + \mathbf{u}(\mathbf{x},t)\cdot\nabla I(\mathbf{x},t) &= 0 \quad \text{in } \Omega\times[0,T], \\ I(\mathbf{x},0) &= I_0(\mathbf{x}) \quad \text{in } \Omega \qquad , \end{aligned} \tag{1}$$

which is refered to as Optical Flow Constraint (OFC) in the literature (Jähne [3]).

This equation is valid under the assumption of constant illumination in the image sequences. Since the brightness in satellite image sequences changes with the altitude of the sun, we will from now on consider a test case without this problem.

Fig. 2 Artificial image sequence $(\mathscr{I}_k)_{k=0}^m := (\mathscr{I}(\mathbf{x}, t_k))_{k=0}^m$: From $t_0 = 0$, $t_{m/2} = 0.5$ to $t_m = 1$

This test case makes it possible to compare our computational results to a known transport field.

We choose the following configuration for the initial image and the transport field:

$$\mathscr{I}_0(\mathbf{x}) = \kappa \sin(\pi x), \quad \mathbf{u}(\mathbf{x}) = (u_1, u_2)^T = (y(y-1), 0)^T,$$

with the computational domain $\Omega = (0,1) \times (0,1)$. This transport field is divergence-free and satisfies the homogeneous Stokes equations.

For this simple setting, we can state an analytic formula for the gray-value distribution

$$\mathscr{I}(\mathbf{x}, t) = \kappa \sin(\pi(x - (ty(y-1)))), \tag{2}$$

which then satisfies the OFC together with the above transport field, the initial image and corresponding periodic boundary data for the image. Choosing discrete time points t_k yields an artificial image sequence $(\mathscr{I}_k)_{k=0}^m$. See Fig. 2 for a visualization of three different time points. We will use this image sequence in the subsequent considerations.

The aim of optical flow estimation is now to reconstruct the transport field $\mathbf{u}(\mathbf{x}, t)$ from the given data of the image sequence. The fact that there is only one equation (see (1)) for two unknown components $\mathbf{u}$ worsen the ill-posed character of the present problem.

The usual way to deal with such a problem is to formulate an optimization problem with the OFC in an appropriate norm as the data term $(\|\partial_t \mathscr{I} + \mathbf{u} \cdot \nabla \mathscr{I}\|_{2,\Omega}^2)$ and supplement this expression with a certain regularization. An example for such a regularization is the enforcement of spatial continuity by the term $\|\nabla \mathbf{u}\|_{2,\Omega}^2$, which leads to the classical Horn and Schunck method [2].

Apart from many different optical flow estimators based on this concept, there exists a tracking-type formulation introduced by Borzí, Ito and Kunisch [1], which avoids the computation of the image derivatives $\partial_t \mathscr{I}$ and $\nabla \mathscr{I}$ from the image data and separates the computational time-step size from the sampling time of the images. The main idea hereby is to state a constrained optimization problem with the OFC as a PDE side condition. Then the time-dependent unknown function $I(\cdot, t)$ is compared to the image sequence at the discrete time points, where an image from the sequence is available. Additional terms for the smoothness of the transport field in space and time are introduced in the cost functional to regularize the underdetermination of the optical flow estimation problem.

We want to set up our solution approach within the second concept, since it yields more promising results than the classical Horn and Schunck method.

3 Physically Based Image Processing and Optimal Control

The sequence of synthetic images of the last section is transported by a solenoidal velocity field, which is chosen as a solution of a homogeneous Stokes system. As a consequence the evolution of the gray-value distribution and the corresponding transport field can be described by the following coupled system:

$$\begin{aligned} \partial_t I(\mathbf{x},t) + \mathbf{u}(\mathbf{x},t)\cdot\nabla I(\mathbf{x},t) &= 0, \\ \partial_t \mathbf{u}(\mathbf{x},t) - \nu\Delta\mathbf{u}(\mathbf{x},t) + \nabla p(\mathbf{x},t) &= 0, \qquad in\ \Omega\times[0,T], \\ \nabla\cdot\mathbf{u}(\mathbf{x},t) &= 0, \end{aligned} \tag{3}$$

with the initial data $(I(\mathbf{x},0),\mathbf{u}(\mathbf{x},0)) = (\mathscr{I}_0(\mathbf{x}),\mathbf{u}^0(\mathbf{x}))$ and appropriate boundary conditions.

The idea of physically based image processing is now to use this forward system as a side conditions in a constrained optimization problem. Currently there are two approaches in this direction presented in the literature.

The first one introduced by Ruhnau and Schnörr [9] considers the identification of a velocity field by using the same data term as in the Horn and Schunck method but with the steady Stokes system as a PDE constraint. The second idea uses a variant of the vorticity-transport formulation of the unsteady Navier-Stokes equation as side condition. This method suggested by Papadakis and Memin [7] uses "pseudo observations" (available measurements or velocity fields obtained by other optical flow estimators) or the OFC in the data term. Both approaches rely on approximations of the OFC in the data term and have therefore the same shortcomings as the Horn and Schunck approach.

As a consequence, we suggest an approach which combines the features of the just presented model-based optimization with the tracking-type optimal control formulation from the last section. Hereby, we use the whole system (3) as a PDE constraint with the data term

$$J_{\mathscr{D}}(I) = \sum_{k=1}^{N} \|I(t_k) - \mathscr{I}_k\|_{2,\Omega}^2. \tag{4}$$

From the modeling point of view the choice of boundary controls seems to be appropriate, since the transport in the domain is driven by the in- and outflow on the boundary.

Of course this formulation also requires regularization. In contrast to the current literature, we will use a L^2-norm regularization for the Dirichlet boundary controls.

We end up with the following constrained minimization problem:
Minimize the cost functional

$$J(I,\mathbf{q}) = J_{\mathscr{D}}(I) + \frac{\alpha}{2}\int_0^T \|\mathbf{q}(t) - \mathbf{q}^\star(t)\|^2_{2,\Gamma} dt \tag{5}$$

subject to an appropriate weak formulation of the system (3), with $\mathbf{u} = \mathbf{q}$ on the boundary Γ.

The work of May, Rannacher and Vexler [5] considers the easier case of a tracking-type problem for the Laplace equation with L^2-Dirichlet controls. In this setting the authors suggest the use of the very weak formulation of the Laplace equation to derive the optimality system which explicitly contains the boundary control variable.

For the Stokes system, we use the same approach relying on the very weak formulation given in the work of Marušíc-Paloka [4] with $H = \{\rho \in H^1(\Omega), (\rho, 1) = 0\}$ and $V = H^2(\Omega) \cap H_0^1(\Omega)$. As a consequence, we obtain the following optimality system for each t in the time interval $[0, 1]$:

1. Primal equations (initial values $(I(\cdot,0), \mathbf{u}(\cdot,0)) = (\mathscr{I}(\cdot,0), \mathbf{u}^0)$):

$$\begin{aligned} (\partial_t I, \psi) + (\mathbf{u}\cdot\nabla I, \psi) &= 0 \quad \forall \psi \in H_0^1(\Omega), \\ (\partial_t \mathbf{u}, \boldsymbol{\varphi}) - \nu(\mathbf{u}, \Delta\boldsymbol{\varphi}) + \nu < \mathbf{q}, \partial_n\boldsymbol{\varphi} >_\Gamma - (p, \nabla\cdot\boldsymbol{\varphi}) &= 0 \quad \forall \boldsymbol{\varphi} \in V, \\ (\mathbf{u}, \nabla\rho) - < \mathbf{q}, \rho\cdot\mathbf{n} >_\Gamma &= 0 \quad \forall \rho \in H. \end{aligned} \tag{6}$$

2. Adjoint equations (initial values $(L(\cdot,T), \mathbf{z}(\cdot,T)) = (0,0)$):

$$\begin{aligned} (\partial_t L, \tilde{\psi}) - (\mathbf{u}\cdot\nabla\tilde{\psi}, L) &= -J_I'(I,\mathbf{q})(\tilde{\psi}) \quad && \forall \tilde{\psi} \in H^1(\Omega), \\ (\partial_t \mathbf{z}, \tilde{\boldsymbol{\varphi}}) + \nu(\Delta\mathbf{z}, \tilde{\boldsymbol{\varphi}}) - (\nabla r, \tilde{\boldsymbol{\varphi}}) &= (I, \nabla(\tilde{\boldsymbol{\varphi}} L)) && \forall \tilde{\boldsymbol{\varphi}} \in L^2(\Omega), \\ (\tilde{\rho}, \nabla\cdot\mathbf{z}) &= 0 && \forall \tilde{\rho} \in H'. \end{aligned} \tag{7}$$

3. Control equations:

$$< \nu\partial_n\mathbf{z} - r\mathbf{n}, \mu >_\Gamma = \alpha < \mathbf{q} - \mathbf{q}^\star, \mu > \quad \forall \mu \in L^2(\Gamma). \tag{8}$$

4 Numerical Method

For the numerical calculations, we apply a Newton-CG method already implemented in the software package RoDoBo [8]. For this purpose, we assume that the forward system in (6) admits a unique solution. By introducing a solution operator S we formulate a reduced cost functional $j(\mathbf{q}) = J(S(\mathbf{q}), \mathbf{q})$. Each iteration step of the Newton method then consists of solving the linear system $\nabla^2 j(\mathbf{q}^k)\delta\mathbf{q} = -\nabla j(\mathbf{q}^k)$

and updating $\mathbf{q}^{k+1} = \mathbf{q}^k + \delta\mathbf{q}$. The directional derivatives of $j(\cdot)$ are computed by solving certain adjoint PDEs using a finite element method in space with pressure and transport stabilization and the implicit Euler scheme in time. The resulting linear system is solved by CG method avoiding the assembling of the Hessian. Details of this methodology can be found in the work of Meidner [6] and the literature cited therein.

5 Numerical Results

First Experiment ($\mathbf{q}^\star = \mathbf{q}^{prior}$)

In the first experiment, we assume that we have a priori knowledge of the transport field on the boundary $\mathbf{q}^{prior}$. This assumption is motivated by the fact that we already gather a transport field either in a preprocessing step or by measurements from real world experiments.

In Fig. 3, we present the numerical result of a calculation with the exact boundary conditions disturbed by a multiplier c of standard Gaussian noise $\mathbf{n}(\mathbf{x})$

$$\mathbf{q}^\star = (y(y-1), 0)^T + c\mathbf{n}(\mathbf{x}). \tag{9}$$

For values of c up to 0.1, we achieve acceptable results of the parameter estimation process (see Fig. 4). Hence, the method will yield good approximations particularly in situations when we have defective information of the transport field on the boundary.

Second Experiment ($\boldsymbol{q}^\star = 0$)

In the second experiment, we choose a more realistic situation, which is closer to a real world application. In general we do not have prior knowledge by measurements in the case of atmospherical flows. As a consequence the only thing we can do is to set $\mathbf{q}^\star$ equal to zero or guessing functions for the in- and outflow from the scenery.

By choosing $\mathbf{q}^\star = 0$, we end up with an approximation shown in the left picture of Fig. 6. Obviously, we have unexpected transport across the upper and lower boundary. To rule out the possibility of numerical artefacts in the corners of the domain, we take a look at the approximation under mesh refinement. Figure 7 shows the second component of the control $\mathbf{q}_{h,k}$, which in fact represents the y component of the transport field on the boundary. Even when we refine the mesh, we observe non-physical behavior of the solution at the outflow corners.

Nevertheless a closer look at the error distribution shows that in principle the method works and yields relatively good approximation. Therefore, we divided the domain into five subregions (see Fig. 6, right). The absolute and relative L^2-errors are shown in Table 1. The left table refers to the five subregions and the right table

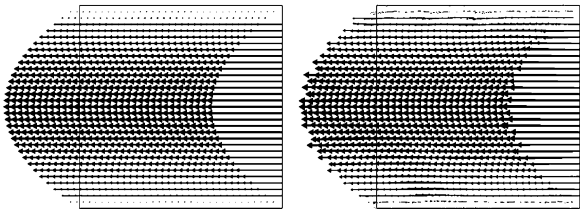


Fig. 3 *Left*: The expected flow field. *Right*: $\mathbf{u}_{h,k}$ for $t = 0.5$ with $\mathbf{q}^\star$ given as the original field disturbed by Gaussian noise $\mathbf{n}(\mathbf{x})$ ($c = 0.04$). L^2-error: $\|\mathbf{u}_{h,k} - \mathbf{u}\|_{L^2(\Omega\times[0,1])} = 3.8892 \cdot 10^{-6}$

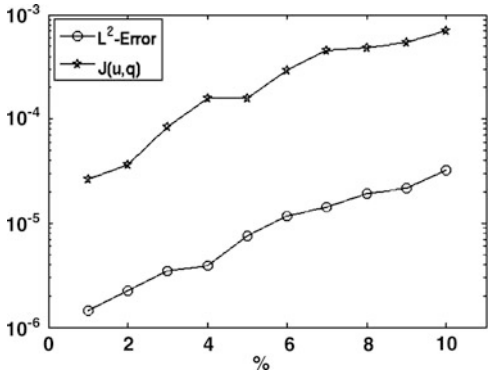
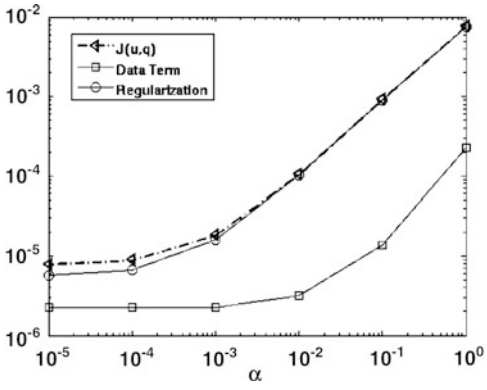


Fig. 4 **First experiment**: Influence of noise on the accuracy of the solution. Up to 10 % standard Gaussian noise the absolute error stays below 10^{-4}

Fig. 5 **Second experiment**: Split-up of the cost functional into the data and the regularization term for a decreasing regularization parameter α



considers further details of the first subregion and documents the shortcomings of the solution in the corners. On the other hand the table also shows that we get a fairly good approximation in the fifth subregion in the interior of the domain.

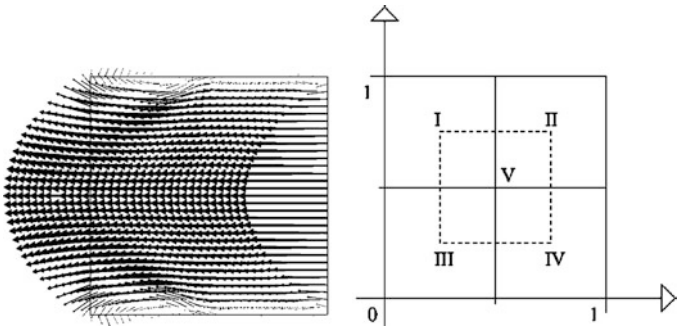


Fig. 6 $\mathbf{q}^\star = 0$. *Left*: Approximation $\mathbf{u}_{h,k}$ for $t = 0.5$. *Right*: Segmentation of the computational domain in five subregions for a further investigation of the error distribution (see Table 1)

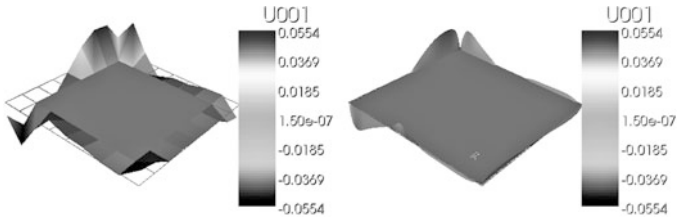


Fig. 7 The y component of the control variable (transport field on the boundary) under mesh refinement. *Left*: 81 nodes. *Right*: 16,641 nodes

Table 1 Detailed error distribution. Absolute and relative errors for eight different subregions. *Left*: I: $\mathbf{x} \in (0, 0.5) \times (0.5, 1)$. II: $\mathbf{x} \in (0.5, 1) \times (0.5, 1)$. III: $\mathbf{x} \in (0, 0.5) \times (0, 0.5)$. IV: $\mathbf{x} \in (0.5, 1) \times (0, 0.5)$. V: $\mathbf{x} \in (0.25, 0.75) \times (0.25, 0.75)$. *Right*: $\mathbf{x} = (x, y) \in (0, 0.5) \times (s, 1)$, with varying s

Subregion	$\|u - u_{h,k}\|_{L^2(\Omega\times[0,1[)}$	Rel. error (%)
I	4.7568e − 5	0.57
II	9.4361e − 6	0.11
III	4.7579e − 5	0.57
IV	9.4375e − 6	0.11
V	6.4346e − 6	0.05

s	$\|u - u_{h,k}\|_{L^2(\Omega\times[0,1])}$	Rel. error (%)
0.95	1.2939e − 5	67.04
0.9	1.8830e − 5	13.20
0.8	3.1905e − 5	3.31

6 Conclusion

We presented an optical flow estimator, which combines a tracking-type optimal control formulation with L^2-Dirichlet boundary control and a coupled evolution model. For known defective data on the boundary the method yields good results, while in the case of lacking prior knowledge the method produces large errors near the boundary corners. A starting point for further investigations could be the usage of a proximal-point like algorithm, which balances the ratio between regularization

and data term. At the moment the regularization dominates the data term in the cost functional (see Fig. 5).

Acknowledgements This article based on collaboration with Prof. Dr. Dr. h.c. R. Rannacher (IAM, University Heidelberg) and Priv.-Doz. Dr. C. S. Garbe (IPM, University Heidelberg). The author gratefully acknowledge both of them for many fruitful discussions and their support.

References

1. Borzí, A., Ito, K., Kunisch, K.: Optimal Control Formulation for Determining Optical Flow. SIAM J. Sci. Comput. **24**(3),818–847 (2002).
2. Horn, B.K.P., Schunck, B.G.: Determining optical flow. Artificial Intelligence. **17**, 185–203 (1981).
3. Jähne, B.: Digital Image Processing. 6th revised and extended Edition, Springer, Heidelberg (2005).
4. Marušíc-Paloka, E.: Solvability of the Navier-Stokes System with L^2 Boundary Data. Appl. Math. Optim. **41** , 365–375 (2000).
5. May, S., Rannacher, R., Vexler, B.: A Priori Error Analysis for the Finite Element Approximation of Elliptic Dirichlet Boundary Control Problems. Proc. ENUMATH-2007, Graz, 637–644 (2007).
6. Meidner, D.: Adaptive Space-Time Finite Element Methods for Optimization Problems Governed by Nonlinear Parabolic Systems. PhD-thesis, Heidelberg University, 2008.
7. Papadakis, N., Memin, E.: Variational Assimilation of Fluid Motion from Image Sequences. SIAM J. on Imag. Sci. (2008).
8. RoDoBo: Software Package for Solving Optimization Problems Governed by PDEs. http://www.rodobo.uni-hd.de
9. Ruhnau, P., Schnörr C.: Optical Stokes Flow Estimation: An Imaging-based Control Approach. Exp. in Fluids. **42**, 61–78 (2007).
10. Schepanski, K., Tegen, I., Macke, A.: Saharan Dust Transport and Deposition Towards the Tropical Northern Atlantic. Atmos. Chem. Phys. **8**, 16061–16096 (2009).

Piecewise Polynomial Collocation for Volterra Integral Equations with Logarithmic Kernels

M. Kolk and A. Pedas

Abstract We propose a numerical method for solving linear Volterra integral equations of the second kind with logarithmic kernels which, in addition to a diagonal singularity, may have a weak boundary singularity. The attainable order of global and local convergence of proposed algorithms is discussed and a collection of numerical results is given.

1 Introduction

Let $\mathbf{R} = (-\infty, \infty)$, $\mathbf{N} = \{1, 2, \ldots\}$, $\mathbf{N}_0 = \{0\} \cup \mathbf{N}$ and let $C^n(\Omega)$ ($n \in \mathbf{N}_0$) be the set of all n times continuously differentiable functions on Ω, $C^0(\Omega) = C(\Omega)$. By $C[0, b]$ we denote the Banach space of continuous functions $z : [0, b] \to \mathbf{R}$ with the norm $\|z\|_\infty = \sup_{0 \le x \le b} |z(x)|$. Let $D_b = \{(x, y) : 0 < y < x \le b\}$ and $\overline{D}_b = \{(x, y) : 0 \le y \le x \le b\}$. Throughout the paper c denotes a positive constant which may have different values by different occurrences.

In many practical applications there arise weakly singular Volterra integral equations

$$u(x) = \int_0^x K(x, y)u(y)dy + f(x), \quad 0 \le x \le b, \tag{1}$$

with $f \in C[0, b]$, $K(x, y) = g(x, y) \log(x - y)$, $g \in C(\overline{D}_b)$. The solutions of such equations are typically non-smooth at the left point of the interval $[0, b]$ where their derivatives become unbounded (see, for example, [2, 10]). In collocation methods the singular behavior of the exact solution can be taken into account by using polynomial splines on special graded grids $\Delta_N^r = \{x_0, \ldots, x_N : 0 = x_0 < \cdots < x_N = b\}$ with the nodes

M. Kolk (✉) · A. Pedas
Institute of Mathematics, University of Tartu, Tartu, Estonia
e-mail: marek.kolk@ut.ee; arvet.pedas@ut.ee

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_22, © Springer-Verlag Berlin Heidelberg 2013

$$x_i = b(i/N)^r, \quad i = 0, \ldots, N, \quad N \in \mathbf{N}, \quad r \in [1, \infty). \tag{2}$$

For $r = 1$ the grid points (2) are uniformly distributed and for $r > 1$ they are more densely situated near the left end point of the interval $[0, b]$ where the solution of (1) may be singular. High-order methods use large values of r (see, e.g., [2,4,9]). However, the use of strongly graded grids by large values of r may cause serious implementation problems since such grids may create unacceptable round-off errors in calculations and therefore lead to unstable behaviour of numerical results.

To avoid problems associated with the use of strongly graded grids the following approach for solving (1) can be used: first we perform in (1) a change of variables so that the singularities of the derivatives of the solution will be milder or disappear and after that we solve the transformed equation by a collocation method on a mildly graded or uniform grid. We refer to [9] for details (see also [1,3,5–8]).

In the present paper we extend the domain of applicability of this approach. To this aim, we examine a more complicated situation for Eq. (1), where the kernel $K(x, y)$, in addition to a logarithmic diagonal singularity, may have a boundary singularity (a singularity as $y \to 0$). Actually, we assume that $K \in W^{n,\lambda}(D_b)$, $n \in \mathbf{N}_0$, $\lambda \in (0, 1)$. Here $W^{n,\lambda}(D_b)$ $(n \in \mathbf{N}_0, \lambda \in (0, 1))$ is the set of functions $K \in C^n(D_b)$ such that, for all $(x, y) \in D_b$ and for all nonnegative integers $i, j \in \mathbf{N}_0$, $i + j \le n$, the following inequalities hold:

$$\left| \left(\frac{\partial}{\partial x} \right)^i \left(\frac{\partial}{\partial x} + \frac{\partial}{\partial y} \right)^j K(x, y) \right| \le c \left\{ \begin{array}{ll} 1 + |\log(x - y)| & \text{if } i = 0 \\ (x - y)^{-i} & \text{if } i > 0 \end{array} \right\} \frac{y^{-\lambda - j}}{1 + |\log y|}. \tag{3}$$

Clearly, $W^{n,\lambda}(D_b) \subset W^{0,\lambda}(D_b)$, $n \in \mathbf{N}$, $\lambda \in (0, 1)$.

The main purpose of the present paper is to extend the corresponding results of [2,4–7,9] to a class of Eq. (1) with kernels $K \in W^{n,\lambda}(D_b)$.

2 Regularity of the Solution

For given $n \in \mathbf{N}$ and $\theta \in (-\infty, 1)$ let $C^{n,\theta}(0, b]$ be the set of functions $u \in C[0, b] \cap C^n(0, b]$ such that

$$|u^{(j)}(x)| \le c \left\{ \begin{array}{ll} 1 & \text{for } j < 1 - \theta \\ 1 + |\log x| & \text{for } j = 1 - \theta \\ x^{1-\theta-j} & \text{for } j > 1 - \theta \end{array} \right\}, \quad x \in (0, b], \quad j = 1, \ldots, n. \tag{4}$$

Clearly, $C^n[0, b] \subset C^{n,\theta}(0, b] \subset C[0, b]$ for any $n \in \mathbf{N}$ and $\theta \in (-\infty, 1)$. It follows from [10] that the regularity of the solution of Eq. (1) can be characterized by the following

Theorem 1. *Let $f \in C[0, b]$, $K \in W^{0,\lambda}(D_b)$, $\lambda \in (0, 1)$. Then Eq. (1) has a unique solution $u \in C[0, b]$. Moreover, if $K \in W^{n,\lambda}(D_b)$, $f \in C^{n,\lambda}(0, b]$, $n \in \mathbf{N}$, $\lambda \in (0, 1)$, then $u \in C^{n,\lambda}(0, b]$.*

3 Smoothing Transformations

For given $\rho \in [1, \infty)$ let $\varphi = \varphi_\rho$ be a transformation of the form

$$\varphi(s) = b^{1-\rho} s^\rho, \quad 0 \le s \le b, \tag{5}$$

or

$$\varphi(s) = 2^{\rho/2} b \left(\sin\left(\frac{\pi}{4b} s \right) \right)^\rho, \quad 0 \le s \le b. \tag{6}$$

Clearly, in both cases, $\varphi \in C^1[0,b]$, $\varphi(0) = 0$, $\varphi(b) = b$ and $\varphi'(s) > 0$ for $0 < s < b$. Thus, φ maps $[0,b]$ onto $[0,b]$ and has a continuous inverse $\varphi^{-1} : [0,b] \to [0,b]$. Note that in the case (5), $\varphi(s) \equiv s$ for $\rho = 1$. We are interested in transformations (5) and (6) with $\rho > 1$ since they possess a smoothing property for $u(\varphi(s))$ with respect to the singularities of $u'(x), \dots, u^{(n)}(x)$ at $x = 0$ (see Lemma 1).

Lemma 1. *[9]. Let $u \in C^{n,\theta}(0,b]$, $n \in \mathbf{N}$, $\theta \in (-\infty, 1)$. Furthermore, let $u_\varphi(s) = u(\varphi(s))$, $s \in [0,b]$, where φ is defined by (5) or (6) with $\rho \in \mathbf{N}$ in case $\rho \le n$ and with $\rho \in \mathbf{R}$ in case $\rho > n$. Then $u_\varphi \in C^{n,1-\rho(1-\theta)}(0,b]$.*

4 Numerical Method

Using (5) and (6) we introduce in (1) the change of variables $y = \varphi(s)$, $x = \varphi(t)$, $s, t \in [0,b]$. We obtain an integral equation of the form

$$u_\varphi(t) = \int_0^t K_\varphi(t,s) u_\varphi(s) ds + f_\varphi(t), \quad 0 \le t \le b, \tag{7}$$

where $f_\varphi(t) = f(\varphi(t))$, $K_\varphi(t,s) = K(\varphi(t), \varphi(s))\varphi'(s)$ are given functions and $u_\varphi(t) = u(\varphi(t))$ is a function which we have to find. Clearly, $f_\varphi \in C[0,b]$ for $f \in C[0,b]$. By (3) with $i = j = 0$ we obtain that $K_\varphi \in C(D_b)$ and

$$|K_\varphi(t,s)| \le c(1 + |\log(t-s)|) s^{-\lambda}, \quad (t,s) \in D_b. \tag{8}$$

For given integers $m, N \in \mathbf{N}$ let $S_{m-1}^{(-1)}(\Delta_N^r) = \left\{ v_N : v_N\big|_{[x_{i-1},x_i]} \in \pi_{m-1}, i = 1, \dots, N \right\}$ be the underlying spline spaces of piecewise polynomial functions on the grid Δ_N^r with the nodes (2). Here $v_N\big|_{[x_{i-1},x_i]}$ $(i = 1, \dots, N)$ is the restriction of $v_N(t)$, $t \in [0,b]$, to the subinterval $[x_{i-1}, x_i] \subset [0,b]$ and π_{m-1} denotes the set of polynomials of degree not exceeding $m-1$. Note that the elements of $S_{m-1}^{(-1)}(\Delta_N^r)$ may have jump discontinuities at the interior knots $x_1, \dots, x_{N-1}$ of the grid Δ_N^r.

In every subinterval $[x_{i-1}, x_i]$ $(i = 1, \dots, N)$ we introduce $m \in \mathbf{N}$ interpolation (collocation) points

$$x_{ij} = x_{i-1} + \eta_j (x_i - x_{i-1}), \quad i = 1, \ldots, N; \ j = 1, \ldots, m, \tag{9}$$

where $0 \le \eta_1 < \ldots < \eta_m \le 1$ and x_i $(i = 0, \ldots, N)$ are defined by the formula (2). We find an approximation $v_N = v_{N,m,r,\varphi}$ to u_φ, the solution of Eq. (7) (under the conditions of Theorems 2 and 3 below the Eqs. (1) and (7) are uniquely solvable), by collocation method from the following conditions:

$$v_N \in S_{m-1}^{(-1)}(\Delta_N^r), \quad N, m \in \mathbf{N}, \ r \ge 1, \tag{10}$$

$$v_N(x_{ij}) = \int_0^{x_{ij}} K_\varphi(x_{ij}, s) v_N(s) ds + f_\varphi(x_{ij}), \quad i = 1, \ldots, N; j = 1, \ldots, m, \tag{11}$$

with x_{ij}, $i = 1, \ldots, N$; $j = 1, \ldots, m$, given by formula (9).

Having determined the approximation v_N for u_φ, we determine an approximation $u_N = u_{N,m,r,\varphi}$ for u, the solution of Eq. (1), setting

$$u_N(x) = v_N(\varphi^{-1}(x)), \quad 0 \le x \le b. \tag{12}$$

Remark 1. The conditions (10) and (11) form a linear system of algebraic equations whose exact form is determined by the choice of a basis in $S_{m-1}^{(-1)}(\Delta_N^r)$. We refer to [9] for a convenient choice of it.

We define an integral operator T_φ by

$$(T_\varphi z)(t) = \int_0^t K_\varphi(t, s) z(s) \, ds, \quad t \in [0, b]. \tag{13}$$

Due to (8) it follows from [10] the following result.

Lemma 2. *Let $K \in W^{0,\lambda}(D_b)$, $\lambda \in (0, 1)$, and let φ be defined by (5) or (6). Then T_φ is compact as an operator from $L^\infty(0, b)$ to $C[0, b]$ (and hence also from $L^\infty(0, b)$ to $L^\infty(0, b)$ and from $C[0, b]$ to $C[0, b]$).*

Theorem 2. *Let $K \in W^{m,\lambda}(D_b)$, $f \in C^{m,\lambda}(0, b]$, $m \in \mathbf{N}$, $\lambda \in (0, 1)$ and let φ be defined by (5) or (6) with $\rho \in \mathbf{N}$ in case $\rho \le m$ and with $\rho \in \mathbf{R}$ in case $\rho > m$. Furthermore, let the interpolation nodes (9) be used.*

Then method (10)–(12) determines for sufficiently large values of N, say $N \ge N_0$, a unique approximation u_N to u, the solution of Eq. (1), and the following error estimate holds:

$$\|u - u_N\|_\infty := \sup_{0 \le x \le b} |u(x) - u_N(x)| \le c \left\{ \begin{array}{ll} N^{-r\varrho(1-\lambda)} & \text{for } 1 \le r < \frac{m}{\varrho(1-\lambda)} \\ N^{-m} & \text{for } r \ge \frac{m}{\varrho(1-\lambda)}, \ r \ge 1 \end{array} \right\}. \tag{14}$$

Here c is a positive constant which is independent of N.

Proof. We consider (7) as an operator equation $u_\varphi = T_\varphi u_\varphi + f_\varphi$ in $L^\infty(0,b)$. Due to Lemma 2 this equation is uniquely solvable and its solution u_φ belongs to $C[0,b]$. Since $u_\varphi(t) = u(\varphi(t))$, $t \in [0,b]$, we obtain by Theorem 1 and Lemma 1 that $u_\varphi \in C^{m,1-\rho(1-\lambda)}(0,b]$. Actually, this result permits an improvement: since $u \in C^{m,\lambda}(0,b]$ and $0 < \lambda < 1$, we will have for u_φ that

$$|u_\varphi^{(j)}(t)| \le c t^{\rho(1-\lambda)-j}, \quad t \in (0,b], \quad j = 1,\dots,m. \tag{15}$$

Further, the conditions (10) and (11) have the operator equation representation $v_N = P_N T_\varphi v_N + P_N f_\varphi$, with T_φ defined by (13), and with an interpolation operator $P_N = P_N^{(m-1)}$ which assigns to every continuous function $z \in C[0,b]$ its piecewise polynomial function $P_N z \in S_{m-1}^{(-1)}(\Delta_N^r)$ such that $P_N z$ interpolates z at the nodes (9): $(P_N z)(x_{ij}) = z(x_{ij})$, $i = 1,\dots,N$, $j = 1,\dots,m$. Due to Lemma 2 we obtain in a similar way as in [5] (see also [9, 11]) that equation $v_N = P_N T_\varphi v_N + P_N f_\varphi$ has for $N \ge N_0$ a unique solution $v_N \in S_{m-1}^{(-1)}(\Delta_N^r)$ and

$$\|u_\varphi - v_N\|_\infty \le c \|u_\varphi - P_N u_\varphi\|_\infty \le c \max_{i=1,\dots,N} \sup_{x \in [x_{i-1},x_i]} \int_x^{x_i} (s-x)^{m-1} |u_\varphi^{(m)}(s)|\, ds, \tag{16}$$

where u_φ is the solution of Eq. (7). Using (15) we get

$$\begin{aligned} &\sup_{x \in [x_{i-1},x_i]} \int_x^{x_i} (s-x)^{m-1} |u_\varphi^{(m)}(s)|\, ds \\ &\qquad \le c\, (x_i - x_{i-1})^m x_i^{\rho(1-\lambda)-m} \le c\, N^{-m} \left(\tfrac{i}{N}\right)^{r\rho(1-\lambda)-m}, \quad i = 1,\dots,N. \end{aligned} \tag{17}$$

Due to (12) we have

$$\sup_{0 \le x \le b} |u(x) - u_N(x)| = \sup_{0 \le t \le b} |u(\varphi(t)) - u_N(\varphi(t))| = \|u_\varphi - v_N\|_\infty.$$

This together with (16) and (17) yields (14). ⋄

Remark 2. Theorem 2 proposes, in particular, how r and ρ should be chosen to achieve the highest convergence order $\|u - u_N\|_\infty = \|v - v_N\|_\infty \le cN^{-m}$ by splines of degree $m-1$. Especially, it follows from Theorem 2 that the accuracy $\|u - u_N\|_\infty \le cN^{-m}$ can be achieved on a mildly graded or uniform grid. As an example, if we assume that $\lambda = 3/4$, $m = 3$ (the case of piecewise quadratic polynomials), $\varrho \ge 12$, the maximal convergence order $\|u - u_N\|_\infty \le cN^{-3}$ is available for $r \ge 1$. In particular, the uniform grid with nodes (2), $r = 1$, may be used. Note also that for $\varphi(s) \equiv s$ (see (5) with $\varrho = 1$) Theorem 2 establishes the order of global convergence of a piecewise polynomial collocation method applied directly (without any change of variables) to the integral equation (1).

In addition to Theorem 2, assuming some additional smoothness of K and f and choosing more carefully the collocation parameters $\eta_1, \ldots, \eta_m$ in (9), in a similar way as in [7], the following superconvergence result can be established.

Theorem 3. *For a given $m \in \mathbf{N}$, assume that $K \in W^{m+1,\lambda}(D_b)$ and $f \in C^{m+1,\lambda}(0,b]$, $\lambda \in (0,1)$. Let φ be defined by (5) or (6) with $\rho \in \mathbf{N}$ in case $\rho \le m+1$ and with $\rho \in \mathbf{R}$ in case $\rho > m+1$. Furthermore, let the interpolation nodes (9) be generated by the node points $\eta_1, \ldots, \eta_m$ of a quadrature approximation*

$$\int_0^1 z(s)\,ds \approx \sum_{j=1}^{m} w_j z(\eta_j), \quad 0 \le \eta_1 < \cdots < \eta_m \le 1, \tag{18}$$

which, with appropriate weights $\{w_j\}$, is exact for all polynomials of degree not exceeding m.

Then method (10)–(12) determines for sufficiently large values of N, say $N \ge N_0 \ge 2$, a unique approximation u_N to u, the solution of Eq. (1), and the following error estimate holds:

$$\gamma_N^{(\rho,r)} := \max_{i=1,\ldots,N;\, j=1,\ldots,m} |u(\varphi(t_{ij})) - u_N(\varphi(t_{ij}))| \le c\, E_N^{(m,\lambda,\rho,r)}. \tag{19}$$

Here c is a positive constant not depending on N and

$$E_N^{(m,\lambda,\rho,r)} = \left\{ \begin{array}{ll} N^{-2r\rho(1-\lambda)} & \text{for } 1 \le r < \frac{m+1}{2\rho(1-\lambda)} \\ N^{-m-1}\log N & \text{for } r \ge \frac{m+1}{2\rho(1-\lambda)}, r \ge 1 \end{array} \right\}. \tag{20}$$

5 Numerical Example

Let us consider the following equation:

$$u(x) = \int_0^x \big(\log(x-y)\big)\, \frac{y^{-8/10}}{1-\log y} u(y)\,dy + 1\,, \quad 0 \le x \le 1. \tag{21}$$

This is an equation of the form (1) where $K(x,y) = \big(\log(x-y)\big)\frac{y^{-8/10}}{1-\log y}$ and $f(x) \equiv 1$. Thus, $K \in W^{m,8/10}(D_1)$ and $f \in C^m[0,1]$ for arbitrary $m \in \mathbf{N}$.

Equation (21) was solved numerically by method (10)–(12) for $\varphi(t) = t^\rho$, $m = 3$, $\eta_1 = (5-\sqrt{15})/10$, $\eta_2 = 1/2$, $\eta_3 = (5+\sqrt{15})/10$. Here η_1, η_2, η_3 are the node points of the Gauss-Legendre quadrature rule by $m = 3$ which is exact for all polynomials of degree not exceeding $2m-1 = 5$.

In Table 1 some results for different values of the parameters N, ϱ and r are presented. In particular, the quantities $\varepsilon_N^{(\varrho,r)}$ in Table 1 are approximate values of the norm $\|u - u_N\|_\infty$, calculated as follows:

Table 1 $\lambda = 8/10$, transformation $\varphi(t) = t^{\rho}$

N	$\varepsilon_N^{(1,1)}$ $\delta_N^{(1,1)}$	$\varepsilon_N^{(7,1)}$ $\delta_N^{(7,1)}$	$\varepsilon_N^{(15,1)}$ $\delta_N^{(15,1)}$	$\varepsilon_N^{(4,4)}$ $\delta_N^{(4,4)}$	$\gamma_N^{(1,1)}$ $\tilde{\delta}_N^{(1,1)}$	$\gamma_N^{(7,1)}$ $\tilde{\delta}_N^{(7,1)}$	$\gamma_N^{(10,1)}$ $\tilde{\delta}_N^{(10,1)}$	$\gamma_N^{(4,4)}$ $\tilde{\delta}_N^{(4,4)}$
16	4.9E−1	2.3E−3	5.7E−4	7.7E−4	5.4E−2	9.1E−6	2.1E−5	6.3E−5
	1.06	**1.71**	**3.67**	**4.18**	**1.00**	**8.17**	**6.54**	**4.95**
32	4.7E−1	8.5E−4	1.5E−4	1.8E−4	5.5E−2	1.1E−6	3.3E−6	1.3E−5
	1.06	**1.65**	**5.58**	**5.71**	**1.03**	**7.02**	**9.90**	**8.11**
64	4.4E−1	3.2E−4	2.8E−5	3.2E−5	5.3E−2	1.6E−7	3.3E−7	1.6E−6
	1.07	**2.63**	**6.77**	**6.79**	**1.05**	**6.93**	**12.22**	**10.86**
128	4.1E−1	1.2E−4	4.1E−6	4.8E−6	5.1E−2	2.3E−8	2.7E−8	1.5E−7
	1.07	**2.62**	**7.41**	**7.41**	**1.08**	**6.90**	**12.49**	**12.48**
256	3.8E−1	4.6E−5	5.5E−7	6.4E−7	4.7E−2	3.3E−9	2.2E−9	1.2E−8
	1.08	**2.62**	**7.73**	**7.72**	**1.10**	**6.89**	**11.99**	**13.04**
512	3.6E−1	1.8E−5	7.2E−8	8.3E−8	4.3E−2	4.8E−10	1.8E−10	8.9E−10
	1.08	**2.62**	**7.89**	**7.89**	**1.12**	**6.90**	**14.56**	**14.47**
1,024	3.3E−1	6.7E−6	9.1E−9	1.1E−8	3.8E−2	7.0E−11	1.2E−11	6.2E−11
	1.15	2.64	8.00	8.00	1.32	6.96	14.40	14.40

$$\varepsilon_N^{(\varrho,r)} = \max_{i=1,\dots,N;\, j=0,\dots,20} |u((\tau_{ij}^{(r)})^{\varrho}) - u_N((\tau_{ij}^{(r)})^{\varrho})|.$$

Here $\tau_{ij}^{(r)} = x_{i-1} + j(x_i - x_{i-1})/20$, $i = 1, \dots, N$, $j = 0, \dots, 20$, with the grid points x_i, defined by formula (2) for $b = 1$, and u_N is the approximate solution for Eq. (21) obtained by method (10)–(12). Since we do not know the exact solution u of Eq. (21), we have used in the role of u the approximation $u_{8,192}$ obtained by the same method with the parameters $m = 3$, $N = 2^{13} = 8{,}192$, $\rho = 8$, $r = 2$ and $\varphi(t) = t^8$:

$$u(x) \approx u_{8,192}(x) = \sum_{j=1}^{3} \beta_{ij} \prod_{k=1, k\neq j}^{3} \frac{x^{1/8} - \eta_k}{\eta_j - \eta_k}, \quad x \in [x_{i-1}, x_i], \quad i = 1, \dots, 8{,}192,$$

where β_{ij} $(i = 1, \dots, 8{,}192;\ j = 1, 2, 3)$ is the solution of a linear system which corresponds to the conditions (10) and (11) (see Remark 1). The ratios

$$\delta_N^{(\varrho,r)} = \varepsilon_{N/2}^{(\varrho,r)} \Big/ \varepsilon_N^{(\varrho,r)}, \quad \tilde{\delta}_N^{(\varrho,r)} = \gamma_{N/2}^{(\varrho,r)} \Big/ \gamma_N^{(\varrho,r)}$$

characterizing the observed convergence rate, are also presented. Here the quantities $\gamma_N^{(\rho,r)}$ are defined in the formula (19).

From Theorems 2 and 3 it follows that for sufficiently large values of N,

$$\delta_N^{(\varrho,r)} = \left\{ \begin{array}{ll} 2^{r\rho/5} \text{ if } & 1 \le r < 15/\rho \\ 2^3 \text{ if } & r \ge 15/\rho \end{array} \right\}, \quad \tilde{\delta}_N^{(\varrho,r)} = \left\{ \begin{array}{ll} 2^{2r\rho/5} \text{ if } & 1 \le r < 10/\rho \\ 2^4 \frac{\log(N/2)}{\log N} \text{ if } & r \ge 10/\rho \end{array} \right\}. \tag{22}$$

In particular, $\delta_N^{(1,1)}$, $\delta_N^{(7,1)}$, $\delta_N^{(15,1)}$, $\delta_N^{(4,4)}$, $\tilde{\delta}_N^{(1,1)}$, $\tilde{\delta}_N^{(7,1)}$, $\tilde{\delta}_N^{(10,1)}$ and $\tilde{\delta}_N^{(4,4)}$ ought to be approximately 1.15, 2.64, 8.00, 8.00, 1.32, 6.96, 14.40 and 14.40, respectively. These values are given in the last row of Table 1.

As we see from Table 1, the obtained numerical results are in agreement with the theoretical estimates.

Acknowledgements This work has been supported by Estonian Science Foundation (grant No. 9104).

References

1. P. Baratella, A. P. Orsi. A new approach to the numerical solution of weakly singular Volterra integral equations. J. Comput. Appl. Math., 163:401–418, 2004.
2. H. Brunner, A. Pedas, G. Vainikko. The piecewise polynomial collocation method for nonlinear weakly singular Volterra eqnations. Math. Comput., 68:1079–1095, 1999.
3. T. Diogo, S. McKee, T. Tang. Collocation methods for second-kind Volterra integral equations with weakly singular kernels. Proc. Roy. Soc. Edinburgh, 124:199–210, 1994.
4. M. Kolk. A Collocation Method for Volterra Integral Equations. Simos, T.E., Numerical Analysis and Applied Mathematics ICNAAM 2010 (CP 1281), Amer. Inst. Physics., Melville, New York, 1187–1190, 2010.
5. M. Kolk, A. Pedas. Numerical solution of Volterra integral equations with weakly singular kernels which may have a boundary singularity. Math. Model. Anal., 14(1):79–89, 2009.
6. M. Kolk, A. Pedas. Numerical Solution of Volterra Integral Equations with Weak Singularities. G. Kreiss, Numerical Mathematics and Advanced Applications 2009, Springer-Verlag Berlin Heidelberg, 507–514, 2010.
7. M. Kolk, A. Pedas, G. Vainikko. High order methods for Volterra integral equations with general weak singularities. Numerical Functional Analysis and Optimization, 30: 1002–1024, 2009.
8. G. Monegato, L. Scuderi. High order methods for weakly singular integral equations with nonsmooth input functions. Math. Comput.,67:1493–1515, 1998.
9. A. Pedas, G. Vainikko. Smoothing transformation and piecewise polynomial collocation for weakly singular Volterra integral equations. Computing, 73:271–293, 2004.
10. A. Pedas, G. Vainikko. Integral equations with diagonal and boundary singularities of the kernel. ZAA, 25(4):487–516, 2006.
11. G. Vainikko. Multidimensional Weakly Singular Integral Equations. Springer- Verlag, Berlin, 1993.

Curvature Calculations for the Level-Set Method

K.Y. Lervåg and Å. Ervik

Abstract The present work illustrates a difficulty with the level-set method to accurately capture the curvature of interfaces in regions that are of equal distance to two or more interfaces. Such regions are characterized by kinks in the level-set function where the derivative is discontinuous. Thus the standard discretization scheme is not suitable. Three discretization schemes are outlined that are shown to perform better than the standard discretization on two selected test cases.

1 Introduction

This article addresses the calculation of interface curvature with the level-set method. In the level-set method, the normal vector and the curvature of an interface can be calculated directly from the level-set function. These calculations are usually done with standard finite-difference methods, typically the second-order central-difference scheme (CD-2) [4, 10, 12].

A problem with these calculations may arise when the level-set function is defined to be a signed-distance function. The signed-distance function is in general not smooth, as can be seen in Fig. 1. Here the derivative of the level-set function

K.Y. Lervåg (✉)
Department of Energy and Process Engineering, Norwegian University of Science and Technology, Kolbjørn Hejes veg 2, NO-7491, Trondheim, Norway
e-mail: karl.yngve@lervag.net

Å. Ervik
SINTEF Energy Research, Sem Sælands veg 11, NO-7465, Trondheim, Norway
e-mail: asmund.ervik@sintef.no

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_23,

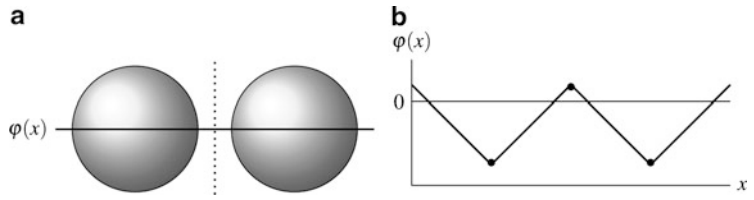


Fig. 1 (**a**) Two droplets in near contact. The *dotted line* marks a region where the derivative of the level-set function is not defined. (**b**) A one-dimensional slice of the level-set function $\varphi(x)$. The *dots* mark points where the derivative of $\varphi(x)$ is not defined

will be discontinuous at the regions that are of equal distance to more than one interface. When two droplets as in Fig. 1 are in near contact, such discontinuities, or kinks, may lead to significant errors when calculating the interface geometries with standard finite difference methods.

2 Governing Equations

2.1 Navier-Stokes Equations for Two-Phase Flow

Consider a domain $\Omega = \Omega^+ \cup \Omega^-$, where Ω^+ and Ω^- denote regions occupied by two respective phases, divided by an interface $\Gamma = \delta\Omega^+ \cap \delta\Omega^-$. The governing equations for incompressible and immiscible two-phase flow in the domain Ω with an interface force on the interface Γ are

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

$$\rho\left(\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u}\cdot\nabla\mathbf{u}\right) = -\nabla p + \nabla\cdot(\mu\nabla\mathbf{u}) + \rho\mathbf{f}_b + \int_\Gamma \sigma\kappa\mathbf{n}\,\delta(\mathbf{x}-\mathbf{x}_I(s))\mathrm{d}s. \tag{2}$$

Here $\mathbf{u}$ is the velocity vector, p is the pressure, $\mathbf{f}_b$ is the specific body force, σ is the coefficient of surface tension, κ is the curvature, $\mathbf{n}$ is the normal unit vector which points into Ω^+, δ is the Dirac delta function, $\mathbf{x}_I(s)$ is a parametrization of the interface, ρ is the density and μ is the viscosity.

It is assumed that the density and viscosity are constant in each phase, but may be discontinuous across the interface. The jump conditions across the interface are

$$[\![\mathbf{u}]\!] = 0, \tag{3}$$

$$[\![p]\!] = 2[\![\mu]\!]\mathbf{n}\cdot\nabla\mathbf{u}\cdot\mathbf{n} + \sigma\kappa, \tag{4}$$

$$[\![\mu\nabla\mathbf{u}]\!] = [\![\mu]\!]\big((\mathbf{n}\cdot\nabla\mathbf{u}\cdot\mathbf{n})\mathbf{nn} + (\mathbf{n}\cdot\nabla\mathbf{u}\cdot\mathbf{t})\mathbf{nt} + (\mathbf{n}\cdot\nabla\mathbf{u}\cdot\mathbf{t})\mathbf{tn} + (\mathbf{t}\cdot\nabla\mathbf{u}\cdot\mathbf{t})\mathbf{tt}\big), \tag{5}$$

where $\mathbf{t}$ is the tangent vector along the interface and $[\![\cdot]\!]$ denotes the jump across an interface, that is $[\![\mu]\!] \equiv \mu^+ - \mu^-$. Note that $\nabla\mathbf{u}$ and (e.g.) $\mathbf{nt}$ are rank-2 tensors. See [3, 4] for more details and a derivation of the interface conditions.

2.2 Level-Set Method

The interface is captured with the zero level set of the level-set function $\varphi(\mathbf{x}, t)$, which is prescribed as a signed-distance function. It is updated by solving an advection equation for φ,

$$\frac{\partial \varphi}{\partial t} + \hat{\mathbf{u}} \cdot \nabla \varphi = 0, \tag{6}$$

where $\hat{\mathbf{u}}$ is the velocity at the interface, extended to the entire domain by solving

$$\frac{\partial \hat{\mathbf{u}}}{\partial \tau} + S(\varphi)\mathbf{n} \cdot \nabla \hat{\mathbf{u}} = 0, \quad \hat{\mathbf{u}}_{\tau=0} = \mathbf{u}, \tag{7}$$

to steady state, cf. [15]. Here τ is a pseudo-time and $S(\varphi) = \varphi/(\varphi^2 + 2\Delta x^2)^{1/2}$ is a smeared sign function which is equal to zero at the interface.

When (6) is solved numerically, the level-set function loses its signed-distance property due to numerical dissipation. The level-set function is therefore reinitialized regularly by solving

$$\begin{aligned} \frac{\partial \varphi}{\partial \tau} + S(\varphi_0)(|\nabla \varphi| - 1) &= 0, \\ \varphi(\mathbf{x}, 0) &= \varphi_0(\mathbf{x}), \end{aligned} \tag{8}$$

to steady state as proposed in [13], where φ_0 is the level-set function that needs to be reinitialized.

Normal vectors and curvatures can be readily calculated from the level-set function as

$$\mathbf{n} = \frac{\nabla \varphi}{|\nabla \varphi|} \quad \text{and} \quad \kappa = \nabla \cdot \left(\frac{\nabla \varphi}{|\nabla \varphi|} \right). \tag{9}$$

3 Numerical Methods

The Navier-Stokes equations (1) and (2) are solved using a projection method on a staggered grid as described in [3, Chap. 5.1.1]. The spatial terms are discretized with CD-2, except for the convective terms which are discretized by a fifth-order WENO scheme. A third-order strong stability-preserving Runge-Kutta (SSP RK) method is

used for the momentum Eq. (2), and a second-order SSP-RK method is used for the level-set Eqs. (6)–(8) [2].

The interface conditions are treated in a sharp fashion with the Ghost-Fluid Method (GFM), which incorporates the discontinuities into the discretization stencils by altering the stencils close to the interfaces, cf. [1, 4, 6]. When using the GFM, the curvature is linearly interpolated from the grid points to the interface before it is used in the discretization stencils for the flow equations unless otherwise stated.

4 Curvature Discretizations

The normal vector and the curvature (9) are typically discretized with the CD-2 at the grid points, cf. [4, 12, 14]. A problem with this is that CD-2 will not converge across kinks, and it may therefore introduce potentially large errors. The errors in the curvature will lead to erroneous pressure jumps at the interfaces, and the errors in the normal vector affect both the discretized interface conditions and the extrapolated velocity (7) which is used in the advection Eq. (6).

A direction difference scheme is presented in [7] which uses a combination of one-sided and central difference schemes to ensure that the differences never cross kinks. The same scheme is used in the present work to calculate the normal vector for MLM and LM (see below). The idea is choose which difference scheme to use based on the values of a quality function,

$$Q(\mathbf{x}) = \left| 1 - |\nabla \varphi(\mathbf{x})| \right| . \tag{10}$$

The quality function is itself calculated with central differences. It effectively detects the regions where the level-set function differs from the signed-distance function. Let $Q_{i,j} = Q(\mathbf{x}_{i,j})$ and $\eta > 0$, then $Q_{i,j} > \eta$ can be used to detect kinks. The parameter η is tuned such that the quality function will detect all the kinks. The value $\eta = 0.1$ is used in the present work.

In the following, three different improved discretization schemes for the curvature are outlined. Note that the first two schemes use the quality function to detect when the improved schemes should be used in favor of CD-2. Also note that the curvature is only calculated at grid points in a narrow band along the interface. At the points where it is not calculated, it is set to zero.

Macklin and Lowengrub's method (MLM) was presented in [8, 9]. With this method, the interface is parametrized with a second-order least-squares polynomial. The curvature is then calculated directly from the parametrization at the desired position on the interface.

To enable easy comparison with the other methods, the estimated curvature values are extrapolated from the interface to the adjacent grid points.

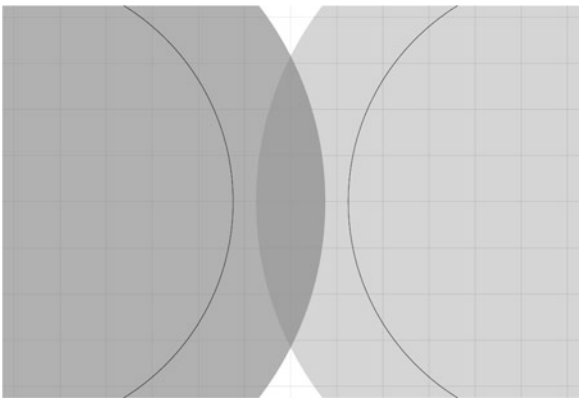


Fig. 2 Simple sketch of how SLM works. The two *circles* are represented by separate level-set functions

Lervåg's method (LM) was presented in [5] and is based on MLM, specifically [8]. The curve parametrization is used to create a local level-set function from which the curvature is calculated on the grid points using CD-2.

The main difference from MLM is that the curvature is calculated at the grid nodes and then interpolated to the interface afterwards. This is argued as a slight simplification of MLM, although an important consequence is that it becomes more important to have an accurate representation of the interface. Instead of using a least-squares parametrization, LM uses monotone cubic Hermite splines.

Salac and Lu's method (SLM) was presented in [11] and is a different approach than MLM and LM. Consider the 2D case of two circles in near contact, see Fig. 2. SLM reconstructs two independent level-set functions ϕ_1 and ϕ_2 for the two circles. The reconstructed functions are then used to calculate the curvature. Since the two reconstructed cones have no kinks, the curvature can be calculated with CD-2. For points close to both circles, a weighted average of the curvature from ϕ_1 and from ϕ_2 is stored. For points close to only one circle, the appropriate curvature is stored. The weighted average is $\kappa = (\kappa_1\phi_2 + \kappa_2\phi_1)/(\phi_1 + \phi_2)$, where the subscripts refer to values calculated on the reconstructed level-set functions. This weighting will prefer κ_1 when closest to circle 1, and vice versa.

5 Comparison of the Discretization Schemes

5.1 A Static Disc Above a Rectangle

Consider a disc of radius r positioned at a distance h above a rectangle, see Fig. 3a. In this case, only the level-set function and the geometrical quantities are considered. None of the governing Eqs. (1), (2) and (6)–(8) are solved.

The parameters used for this case are $r = 0.25$ m and $h = \Delta x$. The domain is 1.5×1.5 m, and the rectangle height is 0.75 m. The grid size is 101×101.

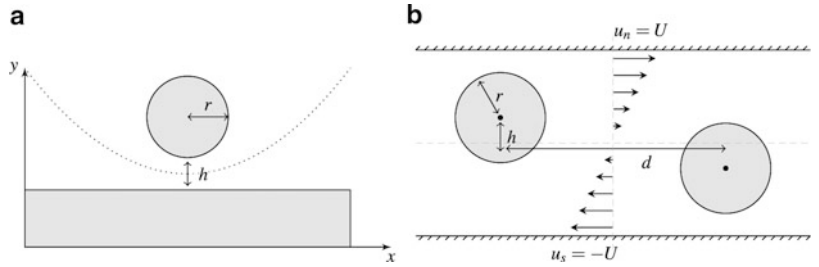


Fig. 3 Initial setup for the *circle* and *rectangle* test, (**a**), and for the drop collision in shear flow test, (**b**). In (**a**), the *dotted line* depicts the kink location, and there is no flow. In (**b**) the flow is indicated by the velocity profile

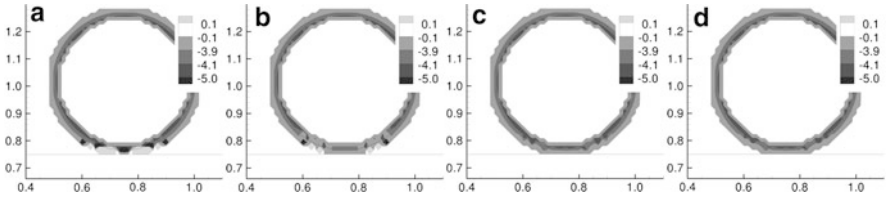


Fig. 4 A comparison of curvature calculations between standard discretization and the improved method. The standard discretization leads to large errors in the curvatures in areas that are close to two interfaces. (**a**) CD-2. (**b**) MLM. (**c**) LM. (**d**) SLM

Figure 4 shows a comparison of the calculated curvatures. The figure shows that CD-2 leads to large errors in the calculated curvatures in the areas that are close to two interfaces. In particular note that the sign of the curvature becomes wrong. The analytic curvature for this case is $\kappa = -1/r = -4$, and the curvature spikes seen for the standard discretization is of the order of $|\kappa| \sim \frac{1}{\Delta x} \simeq 67.3$, which is consistent with the errors seen in [7]. All of the improved methods give much better estimates of the curvature, as expected.

5.2 *Drop Collision in Shear Flow*

Now consider two drops in a shear flow as depicted in Fig. 3b. Both drops have radius r and are initially placed a distance $d = 5r$ apart in the shear flow, where the flow velocity changes linearly from $u_s = -U < 0$ at the bottom wall to $u_n = U$ at the top wall. The computational domain is $12r \times 8r$, and the grid size is 241×161. The density and viscosity differences of the two phases are zero.

The shear flow is defined by the Reynolds number and the Capillary number,

$$Re = \frac{\rho U r}{\mu} \qquad \text{and} \qquad Ca = \frac{\mu U}{\sigma}. \tag{11}$$

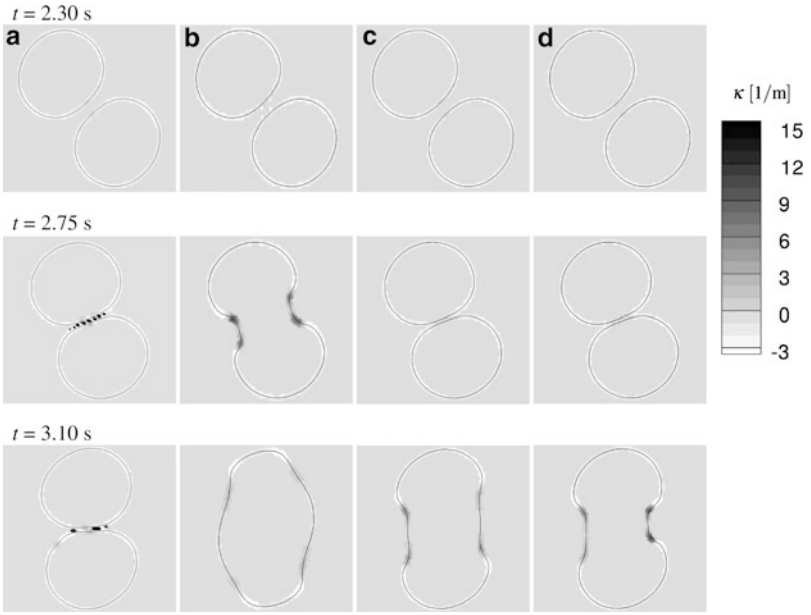


Fig. 5 A comparison between the different discretization schemes of the interface evolution and the curvature κ of drop collision in shear flow. (**a**) CD-2. (**b**) MLM. (**c**) LM. (**d**) SLM

The following results were obtained with $r = 0.5$ m, $h = 0.84r = 0.42$ m, $Re = 10$ and $Ca = 0.025$.

Figure 5 shows a comparison of the interface evolution and the curvature between the different discretization schemes. The first column shows the results with the CD-2. The next three columns show the results with the three improved schemes respectively. The kinks between the drops lead to curvature spikes with CD-2, whereas the improved discretizations calculate the curvature along the kink in a much more reliable manner. LM and SLM give very similar results. This is most likely due to the fact that both these methods calculate the curvature at the grid points and then interpolate, resulting in very similar algorithms as long as the curvature calculations are accurate. MLM on the other hand removes the interpolation step and calculates the curvature directly on the interface. Note that the difference is mainly that the MLM results in slightly earlier coalescence in the given case.

The curvature spikes in obtained with CD-2 are seen to prevent coalescence. This is due to the effect they have on the pressure field as displayed in Fig. 6. Here it is shown that the errors in the curvature with CD-2 lead to an erroneous pressure field between the drops. The distortion of the pressure in the thin-film region leads to a flow into the film region that suppresses coalescence. The corresponding result with LM shows that when the pressure is not distorted, it leads to a flow directed out of the thin-film region.

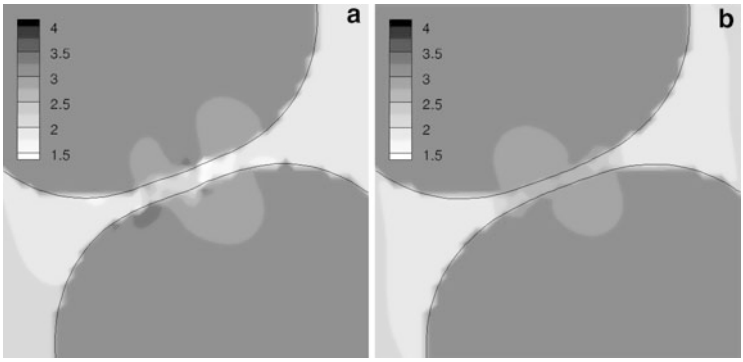


Fig. 6 Comparison of the pressure field in the thin film between the droplets at $t = 2.75$ s. The contour legends indicate the pressure in Pa. (**a**) CD-2. (**b**) LM

6 Conclusions

Three discretization schemes have been implemented to accurately calculate the curvature in regions close to kinks in the level-set function. It has been demonstrated in two test cases that the standard second-order central difference-scheme (CD-2) leads to relatively severe errors across the kinks. Macklin and Lowengrub's method (MLM), Lervåg's method (LM), and Salac and Lu's method (SLM) all give better results. In the second test case where two droplets are put in a shear flow, CD-2 gives a qualitatively different result than all the three improved schemes due to an erroneous pressure field in the thin film region.

Acknowledgements The authors acknowledge Bernhard Müller (Norwegian University of Science and Technology) and Svend Tollak Munkejord (SINTEF Energy Research) for valuable feedback on the manuscript.

This work was financed through the Enabling Low-Emission LNG Systems project, and the authors acknowledge the contributions of GDF SUEZ, Statoil and the Petromaks programme of the Research Council of Norway (193062/S60).

References

1. Fedkiw, R.P., Aslam, T., Merriman, B., Osher, S.: A non-oscillatory Eulerian approach to interfaces in multimaterial flows (the ghost fluid method). Journal of Computational Physics **152**(2), 457–492 (1999). DOI 10.1006/jcph.1999.6236
2. Gottlieb, S., Shu, C.W., Tadmor, E.: Strong stability-preserving high-order time discretization methods. SIAM Review **43**, 89–112 (2001)
3. Hansen, E.B.: Numerical simulation of droplet dynamics in the presence of an electric field. Doctoral thesis, Norwegian University of Science and Technology, Department of Energy and Process Engineering, Trondheim (2005). ISBN 82-471-7318-2

4. Kang, M., Fedkiw, R.P., Liu, X.D.: A boundary condition capturing method for multiphase incompressible flow. Journal of Scientific Computing **15**(3), 323–360 (2000)
5. Lervåg, K.Y.: Calculation of interface curvature with the level-set method. In: Sixth National Conference on Computational Mechanics MekIT'11 (Trondheim, Norway) (23–24 May 2011)
6. Liu, X.D., Fedkiw, R.P., Kang, M.: A boundary condition capturing method for Poisson's equation on irregular domains. Journal of Computational Physics **160**, 151–178 (2000)
7. Macklin, P., Lowengrub, J.: Evolving interfaces via gradients of geometry-dependent interior Poisson problems: Application to tumor growth. Journal of Computational Physics **203**, 191–220 (2005)
8. Macklin, P., Lowengrub, J.: An improved geometry-aware curvature discretziation for level set methods: Application to tumor growth. Journal of Computational Physics **215**, 392–401 (2006)
9. Macklin, P., Lowengrub, J.S.: A new ghost cell/level set method for moving boundary problems: Application to tumor growth. Journal of Scientific Computing **35**, 266–299 (2008)
10. Osher, S., Sethian, J.A.: Fronts propagating with curvature dependent speed: Algorithms based on Hamilton-Jacobi formulations. Journal of Computational Physics **79**, 12–49 (1988)
11. Salac, D., Lu, W.: A local semi-implicit level-set method for interface motion. Journal of Scientific Computing **35**, 330–349 (2008)
12. Sethian, J.A., Smereka, P.: Level set methods for fluid interfaces. Annual Review of Fluid Mechanics **35**, 341–372 (2003)
13. Sussman, M., Smereka, P., Osher, S.: A level set approach for computing solutions to incompressible two-phase flow. Journal of Computational Physics **114**, 146–159 (1994)
14. Xu, J.J., Li, Z., Lowengrub, J., Zhao, H.K.: A level set method for interfacial flows with surfactants. Journal of Computational Physics **212**(2), 590–616 (2006)
15. Zhao, H.K., Chan, T., Merriman, B., Osher, S.: A variational level set approach to multiphase motion. Journal of Computational Physics **127**, 179–195 (1996)

Multimesh $\mathscr{H}_2$-Optimal Model Reduction for Discretized PDEs

S.A. Melchior, V. Legat, and P. Van Dooren

Abstract Model order reduction of a linear time-invariant system consists in approximating its $p \times m$ rational transfer function $H(s)$ of high degree by another $p \times m$ rational transfer function $\widehat{H}(s)$ of much smaller degree. Minimizing the $\mathscr{H}_2$-norm of the approximation error can be achieved iteratively. The convergence behavior of the algorithm depends on the choice of the initial condition. If a large scale dynamical system is obtained by discretizing a partial differential equation on a fine mesh, the efficiency can be improved by taking advantage of several discretizations on coarser meshes. This idea is illustrated on the advection–diffusion equation.

1 Introduction

Computing a small model approximating a large scale dynamical system is a challenging area of research. It is often applied to systems obtained after spatial discretization of a system of *partial differential equations* (PDEs) on a given mesh. Such (un)structured grids are generally composed of a large number of elements in order to model complex geometrical and physical features, and to ensure a good accuracy. Reduced order models are useful to decrease significantly the numerical cost, namely in multiscale simulations to approximate a macroscopic model composed of large micro-scale systems. It can reduce the complexity of control design algorithms for dynamical systems. Moreover, it can be useful for fast time integration in the case of multiple forcing, e.g., through the boundary conditions, or also for allowing larger time steps.

S.A. Melchior (✉) · V. Legat · P. Van Dooren
Université catholique de Louvain (UCL), CESAME, B-1348, Louvain-la-Neuve, Belgium
e-mail: samuel.melchior@uclouvain.be; vincent.legat@uclouvain.be;
paul.vandooren@uclouvain.be

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_24,

In this paper, we focus on the class of implicit linear time-invariant (LTI) systems

$$\begin{bmatrix} M\dot{\mathbf{x}} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{u} \end{bmatrix}, \tag{1}$$

where $M \in R^{N\times N}$, $A \in R^{N\times N}$, $B \in R^{N\times m}$ and $C \in R^{p\times N}$ are the mass, the discrete differential, the input and the output matrices, respectively. Without loss of generality, the matrix $D \in R^{p\times m}$ can be ignored since it is not affected in the model reduction method considered here. The mapping between the input $\mathbf{u}(t) \in R^m$ and the output $\mathbf{y}(t) \in R^p$ of such LTI systems is a convolution. In the Laplace domain, this operator is simply the multiplication by the transfer function $H(s) = C(sM - A)^{-1}B$ of the system; this can be shown by elimination of the state $\mathbf{x}(t) \in R^N$.

The quality of the approximation of a dynamical system by a reduced order model can be measured by comparing their transfer functions. For instance, if a particular norm of their difference is small, it can be shown that the worst case error over the possible trajectories between the outputs of the initial system and its reduced order model is small as well [1].

Assuming $N >> m, p$, the number of dynamic variables can often be reduced, without losing too much accuracy, to a much smaller value n that is independent of the number of elements in the mesh. This is typically achieved by applying a projector $P = V\widehat{M}^{-1}W^T M$ to the state $\mathbf{x}(t)$ which leads to the discrete Petrov-Galerkin projection

$$\underbrace{\begin{bmatrix} s\widehat{M} - \widehat{A} & \widehat{B} \\ \widehat{C} & \end{bmatrix}}_{\widehat{\mathscr{S}}} = \begin{bmatrix} W & \\ & I \end{bmatrix}^T \underbrace{\begin{bmatrix} sM - A & B \\ C & \end{bmatrix}}_{\mathscr{S}} \begin{bmatrix} V & \\ & I \end{bmatrix}. \tag{2}$$

This can be rewritten equivalently in the standard state-space formalism as

$$\underbrace{\begin{bmatrix} sI - \widehat{M}^{-1}\widehat{A} & \widehat{M}^{-1}\widehat{B} \\ \widehat{C} & \end{bmatrix}}_{\widehat{\mathscr{S}}} = \begin{bmatrix} \widehat{M}^{-1}W^T M & \\ & I \end{bmatrix} \underbrace{\begin{bmatrix} sI - M^{-1}A & M^{-1}B \\ C & \end{bmatrix}}_{\mathscr{S}} \begin{bmatrix} V & \\ & I \end{bmatrix}. \tag{3}$$

Then, the reduced order model can be written as

$$\begin{bmatrix} \widehat{M}\dot{\hat{\mathbf{x}}} \\ \hat{\mathbf{y}} \end{bmatrix} = \begin{bmatrix} \widehat{A} & \widehat{B} \\ \widehat{C} & \end{bmatrix} \begin{bmatrix} \hat{\mathbf{x}} \\ \mathbf{u} \end{bmatrix}. \tag{4}$$

2 $\mathscr{H}_2$-Optimal Model Reduction

We are interested in the computation of a reduced order model of the form (4) whose transfer function is denoted by $\widehat{H}(s) = \widehat{C}(s\widehat{M} - \widehat{A})^{-1}\widehat{B}$. One of the most effective model reduction methods is the approximation by balanced truncation. This method guarantees interesting system-theoretical properties, such as stability of the reduced order model and an a priori error bound related to the $\mathscr{H}_\infty$ norm. One major issue in applying this technique to large-scale systems is that the time and storage complexity are $\mathscr{O}(N^3)$ and $\mathscr{O}(N^2)$, respectively. Another approach better suited to discretized boundary value problems is $\mathscr{H}_2$-optimal model reduction, described below.

The error between the transfer functions $E(s) := H(s) - \widehat{H}(s)$ is itself the transfer function of the *error system*

$$\{M_e, A_e, B_e, C_e\} := \left\{ \begin{bmatrix} M & \\ & \widehat{M} \end{bmatrix}, \begin{bmatrix} A & \\ & \widehat{A} \end{bmatrix}, \begin{bmatrix} B \\ \widehat{B} \end{bmatrix}, \begin{bmatrix} C & -\widehat{C} \end{bmatrix} \right\} . \tag{5}$$

The functional to minimize $\mathscr{J} := \|E(\cdot)\|^2_{\mathscr{H}_2}$ can be expressed as:

$$\mathscr{J} = \operatorname{tr}\left(C_e P_e^c C_e^T\right) = \operatorname{tr}((M_e^{-1}B_e)^T M_e^T Q_e M_e (M_e^{-1}B_e)) = \operatorname{tr}\left(B_e^T Q_e B_e\right) , \tag{6}$$

where P_e and $M_e^T Q_e M_e$ are the controllability and observability Gramians of the system (2). They can be obtained from the solutions of the generalized Lyapunov equations:

$$A_e P_e M_e^T + M_e P_e A_e^T + B_e B_e^T = 0 , \quad A_e^T Q_e M_e + M_e^T Q_e A_e + C_e^T C_e = 0 . \tag{7}$$

Using the decompositions $P_e := \begin{bmatrix} P & X \\ X^T & \widehat{P} \end{bmatrix}$ and $Q_e := \begin{bmatrix} Q & Y \\ Y^T & \widehat{Q} \end{bmatrix}$, the following theorem holds.

Theorem 1. *The gradients $\nabla_{\widehat{\mathscr{T}}}\mathscr{J}$ of the error function $\mathscr{J} := \|E(s)\|^2_{\mathscr{H}_2}$ can be expressed as $\begin{bmatrix} \nabla_{\widehat{M}^{-1}\widehat{A}}\mathscr{J} & \nabla_{\widehat{M}^{-1}\widehat{B}}\mathscr{J} \\ \nabla_{\widehat{C}}\mathscr{J} & \end{bmatrix}$ which is given by*

$$2\left(\begin{bmatrix} \widehat{Q}\widehat{M} & \\ & I \end{bmatrix}^T \begin{bmatrix} \widehat{M} & \widehat{B} \\ \widehat{C} & \end{bmatrix} \begin{bmatrix} \widehat{P} & \\ & I \end{bmatrix} + \begin{bmatrix} Y\widehat{M} & \\ & I \end{bmatrix}^T \begin{bmatrix} M & B \\ -C & \end{bmatrix} \begin{bmatrix} X & \\ & I \end{bmatrix} \right) , \tag{8}$$

where

$$\begin{bmatrix} A & \\ & \widehat{A} \end{bmatrix} \begin{bmatrix} X \\ \widehat{P} \end{bmatrix} \widehat{M}^T + \begin{bmatrix} M & \\ & \widehat{M} \end{bmatrix} \begin{bmatrix} X \\ \widehat{P} \end{bmatrix} \widehat{A}^T + \begin{bmatrix} B \\ \widehat{B} \end{bmatrix} \widehat{B}^T = 0 \,, \tag{9}$$

$$\begin{bmatrix} A & \\ & \widehat{A} \end{bmatrix}^T \begin{bmatrix} Y \\ \widehat{Q} \end{bmatrix} \widehat{M} + \begin{bmatrix} M & \\ & \widehat{M} \end{bmatrix}^T \begin{bmatrix} Y \\ \widehat{Q} \end{bmatrix} \widehat{A} + \begin{bmatrix} -C & \widehat{C} \end{bmatrix}^T \widehat{C} = 0 \,. \tag{10}$$

Proof. The proof can easily be derived from Wilson [4].

The gradient forms of Theorem 1 leads to the following key result [4].

Theorem 2. *At every stationary point of* $\mathscr{J}$ *where* $\widehat{P}$ *and* $\widehat{Q}$ *are invertible, Eq. (3) is satisfied with the projection matrices*

$$W := -Y\,\widehat{Q}^{-1}, \quad V := X\,\widehat{P}^{-1} \,, \tag{11}$$

where X, Y, $\widehat{P}$ *and* $\widehat{Q}$ *satisfy the generalized Sylvester equations (9) and (10).*

In addition, it can be shown (see, e.g., [3]) that this choice of projection matrices implies tangential interpolation at the negative of the poles of the reduced order model in directions obtained by multiplying C and B^T by the left and right eigenvectors of $\widehat{M}^{-1}\widehat{A}$, respectively.

Starting with an approximation $\widehat{\mathscr{S}}_k$, the generalized Sylvester equations (9) and (10) can be solved in order to find $(X, Y, \widehat{P}, \widehat{Q})_k = F_1(\widehat{\mathscr{S}}_k)$. Then, the Eq. (11) can be used to compute the new projection matrices $(V, W)_{k+1} = F_2(F_1(\widehat{\mathscr{S}}_k))$. Finally, the Petrov-Galerkin projection (3) leads to the new approximation

$$\widehat{\mathscr{S}}_{k+1} = G(\widehat{\mathscr{S}}_k) := F_3(F_2(F_1(\widehat{\mathscr{S}}_k))) \,. \tag{12}$$

If this iterative scheme converges to a fixed point, Theorem 2 implies that it actually reaches a stationary point of $\mathscr{J}$. Note that the generalized Sylvester equations can be solved efficiently by taking advantage of the sparsity of M and A. For instance, a diagonalization of $\widehat{M}^{-1}\widehat{A}$ uncouples these matrix equations; the solutions of the n resulting linear systems can then be computed with an appropriate solver.

Moreover, the dynamical systems are equivalent under pre- and post-multiplication by invertible matrices. Indeed, it can be verified that, $\forall S, T$ that are nonsingular, the transfer functions corresponding to $\begin{bmatrix} sM - A & B \\ C & \end{bmatrix}$ and $\begin{bmatrix} S & \\ & I \end{bmatrix} \begin{bmatrix} sM - A & B \\ C & \end{bmatrix} \begin{bmatrix} T & \\ & I \end{bmatrix}$ are identical. The resulting quotient structure implies that only the column spaces of V and W matter, but not their basis. In order to choose one unique representative in each class of equivalent systems, the reduced order model can be balanced at each iteration. Nevertheless, this does not affect the convergence of $\mathscr{J}$.

3 Multimesh $\mathscr{H}_2$-Optimal Model Reduction

One important feature of iterative schemes is the need for initial guesses, e.g., $(\widehat{M}, \widehat{A}, \widehat{B}, \widehat{C})$ here. If the latter is sufficiently close to the fixed point, the number of required iterations to reach a given threshold can be reduced compare to other choices of initial conditions. In the case where the dynamical system is derived from PDEs, this leads to two observations. Firstly, the number of state variables is typically large. Hence, the computational cost of each iteration can be very high, especially on fine meshes. Secondly, spatial discretizations on several meshes with a varying range of element sizes can yield a series of dynamical systems of increasing size. As a result, a strategy taking advantage of this additional information could improve the efficiency compared to only performing computations on the finest mesh.

Indeed, a reduced order model computed on the system obtained on a coarse grid can simply be used as initial condition for the fixed-point iterations to compute a reduced order model on a finer grid. No transfer operators are needed between the meshes since a reduced order model is sufficient to start the iteration. For instance, a small number of iterations can be performed on each mesh, starting from the coarsest on which the initial condition could be computed by balanced truncation of the system corresponding to § if its order sufficiently small for this computation.

The proposed strategy leads to the following algorithm:

1. $k = l = 1$ % k and l are the indices of the iteration and the mesh, respectively
2. $(V_k, W_k) \leftarrow$ balanced truncation on (M_l, A_l, B_l, C_l)
3. $\widehat{\mathscr{S}}_k \leftarrow (3)$; $X_k, \widehat{P}_k \leftarrow (9)$; $Y_k, \widehat{Q}_k \leftarrow (10)$; $\nabla_{\widehat{\mathscr{S}}} \mathscr{J} \leftarrow (8)$ using (M_l, A_l, B_l, C_l)
4. IF $\|\nabla_{\widehat{\mathscr{S}}} \mathscr{J}\| < \epsilon$ { $l \leftarrow l + 1$; IF $l > l_{max}$ {STOP} }
5. ELSE { $V_{k+1} = X_k \widehat{P}_k^{-1}$; $W_{k+1} = -Y_k \widehat{Q}_k^{-1}$; $k \leftarrow k + 1$}
6. GOTO 3.

Here l is ranging from 1 to l_{max} for the coarsest and finest ones, respectively. In addition to a user-defined tolerance ϵ, a maximum number of iterations on each mesh could also be imposed in order to limit the computational cost.

4 Numerical Results on the Advection–Diffusion Equation

The unsteady advection–diffusion equation can model the dispersion of a tracer x in a lake in a idealized square domain with an island inside:

$$\frac{\partial x}{\partial t} = -(\underline{w} \cdot \nabla)x + \nabla \cdot (\kappa \nabla x), \text{ in } \Omega, \tag{13}$$

where $\underline{w}$ is the advection velocity and κ is the diffusivity. The domain Ω is a simple geometry obtained by subtracting a small square from a larger one, near its upper

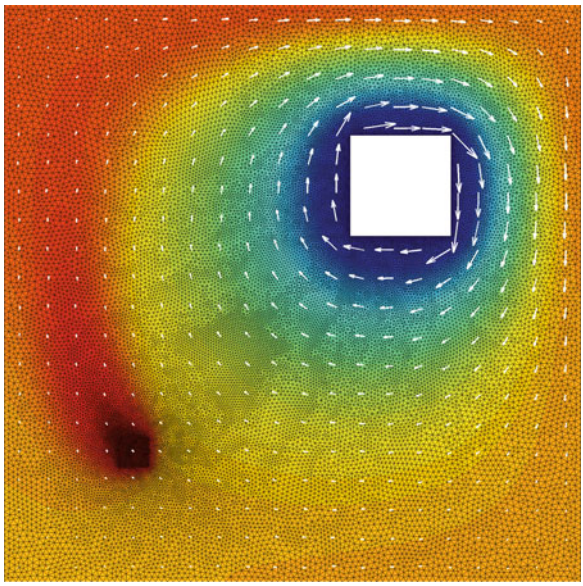


Fig. 1 Steady state solution with $u = 1$ on a mesh with 100 vertices on each outer edge. The *white arrows* show the incompressible velocity field around the squared island

right-hand corner. Homogeneous Neumann and Dirichlet boundary conditions are imposed on the outer boundary and on the inner square in the upper right corner, respectively. The integral of the flux across the latter is the output $y(t) \in R$, while the input $u(t) \in R$ of this SISO system, i.e., $m = p = 1$, is the value imposed as the Dirichlet boundary condition on the smaller inner square in the bottom left corner, which models the source of the tracer.

The first order finite element discretization of this problem on a mesh with N nodes yields a LTI dynamical system of the form (1). The matrix M and A are sparse with most of their lines containing seven nonzero entries, and at most ten otherwise. The Peclet number, defined by $Pe = \frac{WL}{\kappa}$, where W and L are the characteristic velocity and length, is chosen around 1,000. The velocity field around the small square island is incompressible and oriented clockwise. The steady state solution on the finest mesh is shown in Fig. 1.

The fixed-point iterations that converge to the optimal reduced model of degree $n = 3$ in the sense of the $\mathscr{H}_2$ norm is applied to the system obtained from a discretization on a grid with 100 nodes on each outer edge. The convergence of $\mathscr{J}$ and its gradients is shown for the first four iterations in Fig. 2; three different initial conditions are used in order to emphasize the advantage of the multimesh scheme.

The solid lines show the results when $\widehat{\mathscr{T}}_0 = \begin{bmatrix} s-\lambda & & & 1 \\ & s-2\lambda & & 1 \\ & & s-3\lambda & 1 \\ 1 & 1 & 1 & \end{bmatrix}$, where λ is the closest eigenvalue to the imaginary axis. Thus, the first iteration simply consists in the interpolation of the transfer function at $(\lambda, 2\lambda, 3\lambda)$. The choice of $\widehat{B}$

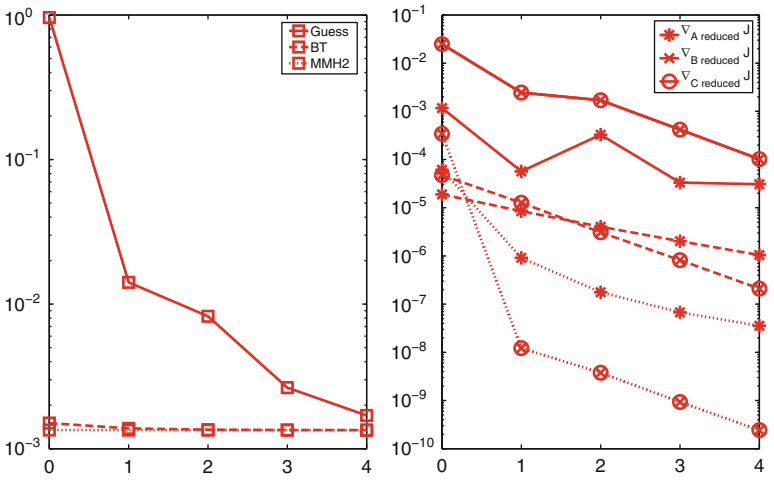


Fig. 2 Convergence of the error function $\mathscr{J}$ (on the *left*) and its gradients with respect to $\widehat{M}^{-1}\widehat{A}$, $\widehat{M}^{-1}\widehat{B}$, and $\widehat{C}$ (on the *right*) with the number of fixed-point iterations. All *curves* are normalized by the $\mathscr{H}_2$ norm of the original dynamical system. Three different initial conditions are used: a simple guess, balanced truncation and the optimal reduced order model in the sense of the $\mathscr{H}_2$ norm on a coarser mesh

and $\widehat{C}$ is not important since a SISO system is considered; both coefficient matrices only modify the value of $\mathscr{J}$ for the initial condition.

The dashed lines and the dotted lines are the results of balanced truncation on the finest mesh and the reduced order model that minimizes the $\mathscr{H}_2$ norm of the approximation error on a coarser mesh, respectively. Those results do not depend on the mesh size, assuming it is sufficiently fine.

This figure confirms that the choice of the initial condition can significantly improve the efficiency. Starting from the simple guess $\widehat{\mathscr{S}}_0$, the approximation becomes rather close to the minimum only after four iterations, while it is already more accurate after the first iteration using one of the two other choices of initial condition. It can be observed that the results of balanced truncation is already almost optimal.

In fact, the error function $\mathscr{J}$ can be written as the sum of $\mathscr{J}_0 = \operatorname{tr}(CPC^T)$, i.e. the $\mathscr{H}_2$ norm of the original dynamical system, and another term that can be computed at the same cost as the gradients. Since the cost of computing $\mathscr{J}_0$ is very high, the convergence can only be inferred using the gradients in practice. It appears that when the coarse mesh is used for the initial condition, the gradients are at least one order of magnitude smaller than with balanced truncation.

Since the number of state variables is very large (above 25,000 on the finest mesh), it is not feasible to compute $\mathscr{J}_0$, λ and the balanced truncation directly. For this figure, they are estimated up to a sufficiently high accuracy using the reduced model of the largest order that minimizes the $\mathscr{H}_2$ norm of the approximation error.

5 Conclusions and Perspectives

The multimesh strategy can yield better convergence results than using balanced truncation at a much cheaper cost. Another advantage is that investigations in order to optimize *order/accuracy* the ratio of the reduced model can be performed on a sufficiently coarse mesh such that the computations are much cheaper than performing one iteration on the finest mesh.

One possible way to improve the numerical conditioning of this algorithm would be to scale the initial system using diagonal balancing based on approximations of the Gramians of the initial system; indeed, it is very expensive to compute the latter exactly. For instance, the approximate Gramians could be obtained as $P \approx \tilde{P} = V\widehat{P}V^T = X\widehat{P}^{-1}X^T$ and $M^TQM \approx M^T\tilde{Q}M = M^TW\widehat{Q}W^TM = M^TY\widehat{Q}^{-1}Y^TM$. Then, the scaling could be chosen such that the diagonal of $\tilde{P}$ and $M^T\tilde{Q}M$ are equal.

The time step for explicit numerical integration must be sufficiently small in order to satisfy a stability condition, such as the CFL condition. The latter depends on the size of the spectrum of the system. The spectrum of the reduced order systems is more clustered. Hence, larger time steps can be chosen.

The most expensive step of the algorithm is to solve for the blocks X and Y in the Sylvester equations (9) and (10). Typically, various discretizations $\mathscr{S}_l$ of the dynamical system are used to solve it iteratively with a multigrid preconditioner. It is interesting to analyze the impact of inexact solves on the convergence of the $\mathscr{H}_2$ iterations.

Using several meshes in order to improve the efficiency of this fixed-point iteration can also be applied on extension to the time-varying case [2] and the nonlinear case which is under development.

Acknowledgements Samuel Melchior is Research fellow with the Belgian National Fund for Scientific Research (FNRS). The present study was carried out within the scope of the project "A second-generation model of the ocean system", which is funded by the Communauté Française de Belgique, as Actions de Recherche Concertées, under Contract ARC 04/09-316. 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 State, Science Policy Office. The scientific responsibility rests with its author(s).

References

1. A. Antoulas. Approximation of Large-Scale Dynamical Systems. SIAM Publications, Philadelphia (2005) SIAM J. Matr. Anal. Appl., Vol.31(5), 2738–2753, 2010.
2. S.A. Melchior, P. Van Dooren, K.A. Gallivan. Model reduction of linear time-varying systems over finite horizons. Submitted to Applied Numerical Mathematics.
3. P. Van Dooren, K.A. Gallivan, P.-A. Absil. $\mathscr{H}_2$-optimal model reduction with higher order poles.
4. D.A. Wilson. Optimum solution of model-reduction problem. In *Proc. IEEE*, **117**:1161–1165, 1970S

The Computation of Long Time Hamiltonian Trajectories for Molecular Systems via Global Geodesics

H. Schwetlick and J. Zimmer

Abstract A string method for the computation of Hamiltonian trajectories linking two given points is presented, based on the Maupertuis principle; trajectories then correspond to geodesics. For local geodesics, convergence of an algorithm based on Birkhoff's method has been shown recently in Schwetlick and Zimmer (Submitted). We demonstrate how to extend this approach to global geodesics and thus arbitrary boundary values of the corresponding Hamiltonian problem. Numerical illustrations of the algorithm are given, as well as situations are shown in which the method converges to a degenerate solution.

1 Introduction

We consider a numerical approximation to the Hamiltonian boundary value problem with kinetic energy $K := \int \frac{1}{2}\dot{q}^2 \,\mathrm{d}x$ and potential energy $P := \int V(q)\,\mathrm{d}x$. Here $q \in Q \subset \mathbb{R}^n$. This class of dynamical systems covers a wide variety of applications, notably the Hamiltonian description of reaction trajectories between two conformational molecular states. The equations of motion are

$$\frac{d^2q(t)}{\mathrm{d}t^2} = -\nabla V(q), \tag{1}$$

supplied with boundary data

$$q(0) = q_a \text{ and } q(T_0) = q_b \tag{2}$$

for given $q_a, q_b \in Q$ and indeterminate $T_0 > 0$.

H. Schwetlick (✉) · J. Zimmer
Department of Mathematical Sciences, University of Bath, Bath, BA2 7AY, UK
e-mail: schwetlick@maths.bath.ac.uk; zimmer@maths.bath.ac.uk

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_25,

The numerical approximation of (1) is an area of great current interest, with two main lines of research being adaptive geometric integrators and variational integrators; we only refer the reader to, e.g., [1,3,5,6,8].

A particular focus of variational integrators is the conservation of energy. We present here a variational approach where energy conservation is automatically built in. Specifically, a classic result is that trajectories of (1) with prescribed total energy E ($E = K + P$) can be found as geodesics, where the metric is the *Jacobi metric*

$$g_{ij}^{\mathrm{Jac}}(q) := (E - V(q))\delta_{ij}(q), \tag{3}$$

with δ_{ij} being the Kronecker delta. We recall that geodesics are critical points of the *length functional*

$$L[\gamma] := \int_a^b \sqrt{2(E - V(q))\,\langle q', q'\rangle}\,\mathrm{d}\tau. \tag{4}$$

The parametrisation

$$t = \int_0^\tau \sqrt{\frac{\langle q', q'\rangle}{2(E - V)}}\,\mathrm{d}s \tag{5}$$

yields then the physical time for a solution of (1).

2 Approximation by a Discrete Birkhoff Procedure

Geodesics as critical points of (4) can be found in a constructive manner by the Birkhoff curve shortening procedure. However, this procedure assumes that sufficiently short geodesics can be computed exactly. A numerical procedure thus needs to ensure that discretisation errors do not lead to erroneous results. A consistent algorithm for local geodesics was given in [11]; here locality of a geodesic means in particular that the geodesic can be represented as a graph. For given initial and final state q_a and q_b, this can be achieved by choosing the total energy E large enough. However, in practice E is given by the physics or chemistry of the problem under consideration as well. An iterative procedure to choose E is presented in Sect. 2.4.

2.1 Existence of Trajectories

For given initial and final point q_a and q_b, there is not always a physical trajectory of (1) linking these points. We now discuss how this classic fact materialises in the Maupertuis setting (that is, geodesics for the Jacobi metric (3)). The configuration space is the set $Q := \{q \in \mathbb{R}^n \mid V(q) < E\}$. This set can be disconnected,

which happens in particular for the case of small total energy E. Also, Q is open. It is possible that geodesics exist only in the closure $\bar{Q}$; am example of a geodesic touching the boundary ∂Q is shown in Sect. 3. Existence results are available for geodesics in Q [2], but the assumptions are formulated in terms of the Riemannian tensor and do not translate easily into estimates in terms of the total and potential energy. A local existence result, with a constructive proof, is given in [11]. We now sketch how to turn this local existence result in a global one. To do so, we first recall the Birkhoff procedure for the continuous (that is, approximation-free) case.

Birkhoff's method assumes that one can join the given points q_a and q_b by a curve γ in Q. Mark $2k+1$ points on the curve so that $q_a = q_0$, $q_i = \gamma(t_i)$ with $t_i < t_{i+1}$ and $q_{2k} = q_b$; here k has to be large enough so that neighbouring points q_i and q_{i+1} are close enough together for $i = 0, \ldots, 2k-1$. Then joint the points with even index by a geodesic (Birkhoff's method assumes that this local geodesic can be found). This defines a new curve γ_1. Identify the points with odd index on that curve by their parametrisation in t. Join now neighbouring points with odd index q_{2i+1} and q_{2i+3} with a local geodesic. This defines a curve γ_2. The length reduces in each of these two steps unless the input already was a piecewise geodesic. One can show that this procedure converges under mild assumptions to a geodesic [4].

2.2 The Algorithm

We assume in this section that the total energy E, the initial and final points q_a and q_b and a curve $\gamma \subset Q$ joining q_a and q_b are given. As discussed above, this is a nontrivial assumption, as E determines Q. However, this assumption can be met by joining q_a and q_b by a straight line; this line will be in Q provided E is large enough. In Sect. 2.4, we describe an iterative procedure to reduce the total energy E.

1. Given this initial curve in Q, mark points $q_a = \gamma(t_0), \gamma(t_1), \ldots, \gamma(t_m) < \gamma(t_{m+1}) = q_b$, for example equidistributed. These points define a polygonal curve, and neighbouring points have to be close enough together. Specifically, consider each triplet of neighbouring points (marked by dashed ellipses in Fig. 1 for two representative triplets). If the points are close enough together, then this triplet can be represented as a graph, even if this is not true for the entire curve. We require that the first triplet (the one including q_a) satisfies the assumptions of the local algorithm described in [11].
2. Then apply the local Birkhoff method of [11] to this triplet. Here the graph structure of this triplet can be taken to be with respect to the line joining the two outer points (dotted line in Fig. 1). The output of the local Birkhoff method is indicated as dashed-dotted line in Fig. 1. We mark the new mid point (e.g., the mid point by arc length parametrisation, which would correspond to the image of the mid point on the straight line under the local Birkhoff procedure; these two points are both marked with dot in Fig. 1).
3. Repeat the previous step for the next triplet (again marked as a dashed ellipsis in Fig. 1). As leftmost point, one could take the corresponding point on the initial

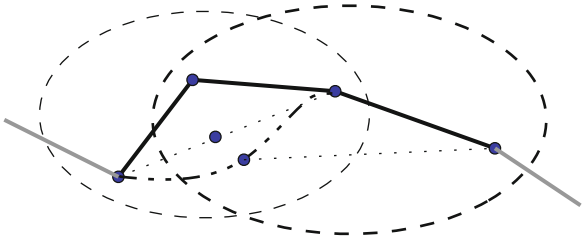


Fig. 1 The geometry of the global Birkhoff algorithm as presented in Sect. 2.2

curve, or the rightmost point obtained in the previous step. In the simulations in Sect. 3, we use the latter. The corresponding straight line defining the local graph structure is indicated as a dashed segment in Fig. 1. Iterate until all triplets are treated.

4. Restart the procedure with the output obtained in the last step, and keep iterating.

2.3 Potential Pitfalls, and Adaptivity as Partial Remedy

In Sect. 3, we present examples where an implementation of this algorithm provides good approximations to long-time trajectories for the Müller potential. However, we point out that, unlike for the local procedure [11], no convergence proof is available. We explain why a general convergence result cannot be expected, even if there is a solution geodesic. The local convergence proof of [11] uses centrally that the end points are fixed and only interior points move. For the global algorithm, this is only true for q_a and q_b but not the end points of the triplets in between. It is then possible that the end points of these inner triplets move to a place where the assumptions made on the triplets no longer hold. For example, the core assumption of the local argument [11] is that there is an invariant neighbourhood of a certain size within the configuration set Q. This is not the case if the triplet slides towards the boundary ∂Q under iteration.

Even if no problem with the boundary occurs, the algorithm may still not converge as stated. Namely, it relies on the ability to compute local (small) geodesics to good accuracy, and thus in particular assumes the existence of local geodesics. This is measured by the radius of injectivity: the local existence of a geodesic emanating from a given point q with a given velocity v follows from standard existence and uniqueness ODE results. The radius of injectivity of a point $q \in Q$ is the maximal radius for which geodesics emanating from this point with arbitrary velocity exist. Since the local convergence argument [11] works with a fixed geometric configuration, the radius of injectivity is bounded from below for the points considered there. If the configuration moves, as the triplets of the global algorithm do, assumptions on the lower bounds on the injectivity radius can be violated.

A possible cure for the latter problem is to introduce additional points so that next to nearest neighbours are in the radius of injectivity of each other.

2.4 The Choice of the Total Energy and the Initial Curve

Typically, the total energy E will be prescribed at least to some extent by the problem under consideration. For given E, it is then not trivial, and sometimes impossible, to join q_a and q_b in Q by a curve. We describe a feasible though computationally expensive approach to find an initial curve.

Suppose the total energy aimed for is E_0. Then join q_a and q_b by a curve such that the curve is contained in the configuration manifold Q_1, say, of a total energy $E_1 > E_0$. A possible but expensive choice is to make E_0 so large that the line segment between q_a and q_b is in Q_1. Apply the algorithm of Sect. 2.2 to obtain the final polygon for this energy level, say γ_1. Then choose $E_2 < E_1$ such that $\gamma_1 \subset Q_2 = \{q \in \mathbb{R}^n \mid V(q) < E_2\}$. Restart the algorithm with E_2 and γ_1 and iterate. We remark that this is a way to determine the appropriate total energy to link two given points if no initial guess based on knowledge or intuition on the underlying problem is available.

3 Numerical Illustrations

As in [12], we consider the model problem of reaction trajectories for the Müller potential [7]. This potential serves as a common nontrivial test case for transition path methods (see, e.g., [9, 10]). We briefly describe the potential, see [12]. The Müller potential has two degrees of freedom; it exhibits three minima, see Fig. 2, with the global minimum of -146.7 located at $(-0.558, 1.442)$. The analytic form of the Müller potential is

$$V(x, y) = \sum_{j=1}^{4} A_j \exp\left[a_j (x - x_j)^2 + b_j (x - x_j)(y - y_j) + c_j (y - y_j)^2\right] \quad (6)$$

with

$$\begin{aligned} A &= (-200, -100, -170, 15), & a &= (-1, -1, -6.5, 0.7), \\ b &= (0, 0, 11, 0.6), & c &= (-10, -10, -6.5, 0.7) \\ x_j &= (1, 0, -0.5, -1), & y_j &= (0, 0.5, 1.5, 1). \end{aligned}$$

Despite the simplicity of this expression, the reaction coordinates are highly contorted, see Fig. 2 for a surface plot.

We now discuss two simulations with this potential. The total energy is chosen to be $E = 20$. First, the points to be connected are $q_a = (-1, 1)$ and $q_b = (-0.05, 0.48)$, a point near the middle minimum. The initial and final curve of a simulation with this potential are shown in Fig. 3. Intermediate steps of the corresponding Birkhoff procedure are shown in Fig. 4. The entire network of passed

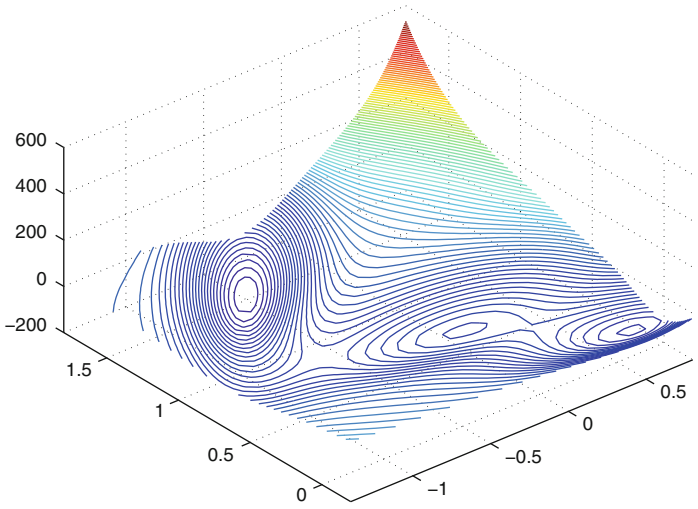


Fig. 2 A plot of the energy surface of the Müller potential

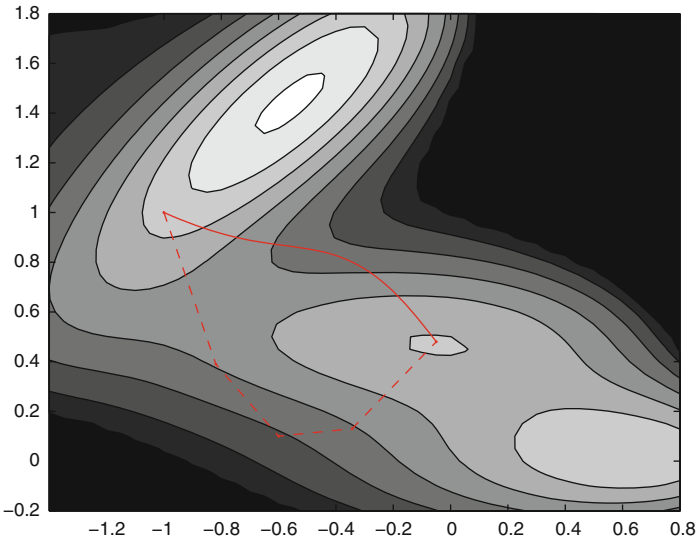


Fig. 3 The initial curve of a simulation (*dashed*), and the corresponding final curve (*solid*)

geodesic connections (light lines in Fig. 4) is computed in a few seconds with MATLAB on a standard PC. Note that the convergence is fast, sweeping over a significant part of the configuration manifold.

In a second simulation, the initial point q_a is moved to the vicinity of the leftmost minimum of the Müller potential, namely $q_a = (-0.6, 1.4)$. In this case the total energy $E = 20$ is so small that the Birkhoff iteration for the initial curve, dashed in Fig. 5, initially produces local geodesic connections which touch the

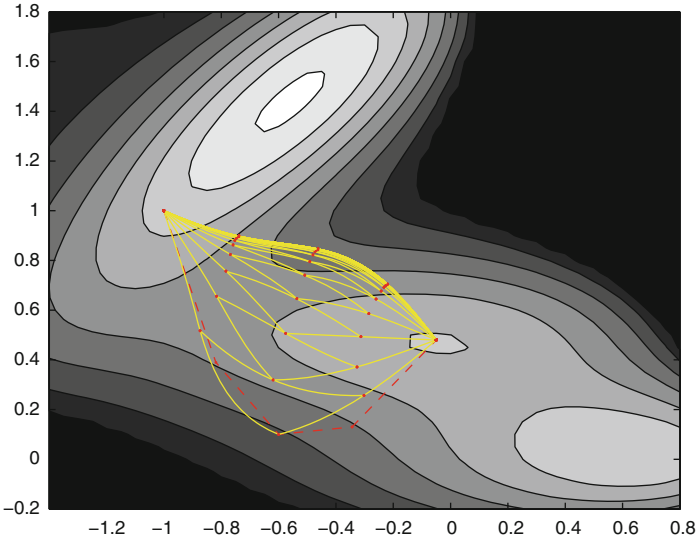


Fig. 4 An illustration of the Birkhoff procedure described in this article, starting from the *dashed line* as initial polygon. The *light lines* indicate the intermediate local geodesic connections. The algorithm terminates if the average movement of the nodes is less than a prescribed tolerance, here $7.5806 \cdot 10^{-7}$

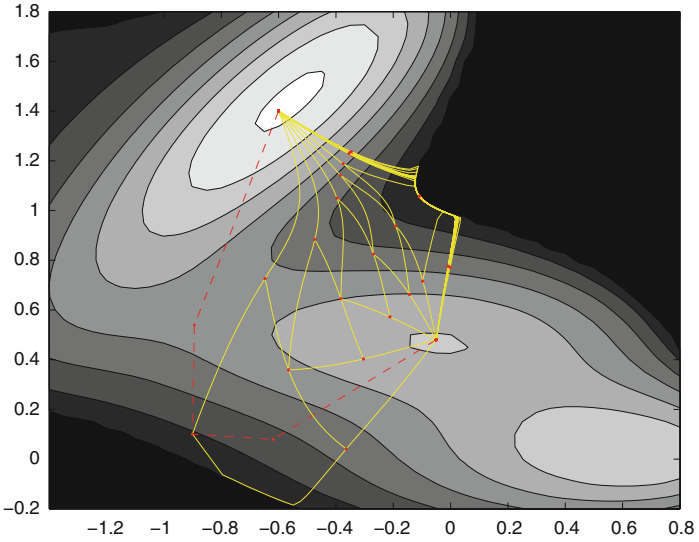


Fig. 5 A further illustration of the Birkhoff procedure for a different initial curve (*dark, dashed*), showing convergence to a degenerate solution. The lower boundary is touched once at the beginning of the iteration (lowest light segment)

boundary ∂Q, where $E = V$, but then move back to the interior of the configuration manifold. At later stages of the iteration, geodesic connections touch the opposite side of the boundary and converge to a degenerate solution. This solution does not correspond to a physical trajectory, since the time reparametrisation (5) diverges as one approaches the boundary ∂Q from the interior.

Even if the simulation fails in the sense that no physical trajectory is found for the given value of E, the fact that the simulation sweeps from one side of the boundary to the other suggests that there exists no interior minimal geodesic. However, it is natural to expect the possibility of a saddle point extremum.

References

1. S. Blanes and C. J. Budd. Adaptive geometric integrators for Hamiltonian problems with approximate scale invariance. *SIAM J. Sci. Comput.*, 26(4):1089–1113 (electronic), 2005.
2. William B. Gordon. The existence of geodesics joining two given points. *J. Differential Geometry*, 9:443–450, 1974.
3. Ernst Hairer, Christian Lubich, and Gerhard Wanner. *Geometric numerical integration*, volume 31 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2002. Structure-preserving algorithms for ordinary differential equations.
4. Jürgen Jost. *Riemannian geometry and geometric analysis*. Universitext. Springer-Verlag, Berlin, third edition, 2002.
5. A. Lew, J. E. Marsden, M. Ortiz, and M. West. Variational time integrators. *Internat. J. Numer. Methods Engrg.*, 60(1):153–212, 2004.
6. J. E. Marsden and M. West. Discrete mechanics and variational integrators. *Acta Numer.*, 10:357–514, 2001.
7. Klaus Müller. Reaction paths on multidimensional energy hypersurfaces. *Angew. Chem. Int. Edit.*, 19(1):1–13, 1980.
8. Sujit Nair. Time adaptive variational integrators: A space-time geodesic approach. *Physica D: Nonlinear Phenomena*, 241(4):315–325, 2012.
9. Roberto Olender and Ron Elber. Calculation of classical trajectories with a very large time step: Formalism and numerical examples. *J. Chem. Phys.*, 105(20):9299–9315, 1996.
10. Daniele Passerone and Michele Parrinello. Action-derived molecular dynamics in the study of rare events. *Phys. Rev. Lett.*, 87(10):108302, Aug 2001.
11. Hartmut Schwetlick and Johannes Zimmer. A convergent string method: Existence and approximation for the Hamiltonian boundary-value problem. Submitted.
12. Hartmut Schwetlick and Johannes Zimmer. Calculation of long time classical trajectories: Algorithmic treatment and applications for molecular systems. *J. Chem. Phys.*, 130(12):124106, 2009.

A Nonlinear Local Projection Stabilization for Convection-Diffusion-Reaction Equations

G.R. Barrenechea, V. John, and P. Knobloch

Abstract We propose a new local projection stabilization (LPS) finite element method for convection-diffusion-reaction equations. The discretization contains a crosswind diffusion term which depends on the unknown discrete solution in a nonlinear way. Consequently, the resulting method is nonlinear. Solvability of the nonlinear problem is established and an a priori error estimate in the LPS norm is proved. Numerical results show that the nonlinear crosswind diffusion term leads to a reduction of spurious oscillations.

1 Introduction

Let $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$, be a bounded polygonal (polyhedral) domain with a Lipschitz-continuous boundary $\partial\Omega$ and let us consider the steady-state convection-diffusion-reaction equation

G.R. Barrenechea
Department of Mathematics and Statistics, University of Strathclyde, 26 Richmond Street, Glasgow, G1 1XH, Scotland
e-mail: gabriel.barrenechea@strath.ac.uk

V. John
Weierstrass Institute for Applied Analysis and Stochastics (WIAS), Mohrenstr. 39, 10117, Berlin, Germany

Department of Mathematics and Computer Science, Free University of Berlin, Arnimallee 6, 14195, Berlin, Germany
e-mail: volker.john@wias-berlin.de

P. Knobloch (✉)
Department of Numerical Mathematics, Faculty of Mathematics and Physics, Charles University, Sokolovská 83, 18675, Praha 8, Czech Republic
e-mail: knobloch@karlin.mff.cuni.cz

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_26,

$$-\varepsilon\,\Delta u + \mathbf{b}\cdot\nabla u + c\,u = f \quad \text{in } \Omega\,, \qquad\qquad u = u_b \quad \text{on } \partial\Omega\,. \tag{1}$$

It is assumed that ε is a positive constant and $\mathbf{b} \in W^{1,\infty}(\Omega)^d$, $c \in L^\infty(\Omega)$, $f \in L^2(\Omega)$, and $u_b \in H^{1/2}(\partial\Omega)$ are given functions satisfying

$$\sigma := c - \frac{1}{2}\nabla\cdot\mathbf{b} \geq \sigma_0 > 0 \quad \text{in } \Omega\,,$$

where σ_0 is a constant. Then the boundary value problem (1) has a unique solution in $H^1(\Omega)$.

The numerical solution of (1) is still a challenge if convection dominates diffusion. In the framework of the finite element method, the common approach is to apply a stabilized method, see [5] for a review. Linear stabilized methods typically provide approximate solutions that possess spurious oscillations in layer regions. These oscillations can be suppressed without smearing the layers significantly by adding an additional artificial diffusion term depending on the approximate solution in a nonlinear way, see [2] for a review of various approaches of this type that we call spurious oscillations at layers diminishing (SOLD) methods.

Here we concentrate on local projection stabilizations (LPS) [1, 3, 4]. In comparison with residual-based methods, the linear LPS has several advantages. In particular, it does not contain second order derivatives, which may be costly to implement, and if applied to systems of PDEs, it does not lead to additional couplings between various unknowns. To suppress oscillations in layer regions, we design a new nonlinear stabilization term inspired by both the linear LPS and the above-mentioned nonlinear SOLD methods. Since we assume that the linear LPS adds enough artificial diffusion in the streamline direction, we introduce only crosswind diffusion through the nonlinear term. To preserve the above-mentioned advantages of the LPS, the residual usually appearing in SOLD terms is replaced by a fluctuation of the crosswind derivative of the approximate solution. This makes sense since the additional stabilization should be added in regions where oscillations in the crosswind direction appear. For the resulting nonlinear method, we prove the existence of a solution, without any restriction on the multiplicative factor in the nonlinear term. Furthermore, we establish an a priori error estimate with respect to the standard LPS norm. The properties of the new method are illustrated by numerical results. Let us mention that such results cannot be obtained using a linear crosswind-diffusion term since then a reduction of spurious oscillations would be possible only at the price of a considerable smearing of the layers.

The plan of the paper is as follows. Section 2 will summarize the main abstract hypothesis imposed on the different partitions of Ω and the finite element spaces considered. Section 3 presents the method whose well-posedness is analyzed in Sect. 4. An a priori error estimate is derived in Sect. 5. Finally, numerical results are presented in Sect. 6.

2 Assumptions

Given $h > 0$, let $W_h \subset W^{1,\infty}(\Omega)$ be a finite-dimensional space approximating the space $H^1(\Omega)$ and set $V_h = W_h \cap H_0^1(\Omega)$. Next, let $\mathscr{M}_h$ be a set consisting of a finite number of open subsets M of Ω such that $\overline{\Omega} = \cup_{M \in \mathscr{M}_h} \overline{M}$. It will be supposed that, for any $M \in \mathscr{M}_h$,

$$\text{card}\{M' \in \mathscr{M}_h\,;\ M \cap M' \neq \emptyset\} \leq C\,, \tag{2}$$

$$h_M := \text{diam}(M) \leq C\,h\,, \tag{3}$$

$$h_M \leq C\,h_{M'} \qquad \forall\ M' \in \mathscr{M}_h,\ M \cap M' \neq \emptyset\,. \tag{4}$$

The space W_h is assumed to satisfy the inverse inequality $|v_h|_{1,M} \leq C\,h_M^{-1}\,\|v_h\|_{0,M}$ for any $v_h \in W_h$, $M \in \mathscr{M}_h$. For any $M \in \mathscr{M}_h$, a finite-dimensional space $D_M \subset L^\infty(M)$ is introduced. It is assumed that there exists a positive constant β_{LP} independent of h such that

$$\sup_{v \in V_M} \frac{(v,q)_M}{\|v\|_{0,M}} \geq \beta_{LP}\,\|q\|_{0,M} \qquad \forall\ q \in D_M,\ M \in \mathscr{M}_h\,,$$

where $V_M = \{v_h \in V_h\,;\ v_h = 0 \text{ in } \Omega \setminus M\}$. Furthermore, for any $M \in \mathscr{M}_h$, a finite-dimensional space $G_M \subset L^\infty(M)$ containing the space D_M is introduced such that $(\partial v_h/\partial x_i)|_M \in G_M$ for any $v_h \in W_h$, $i = 1, \ldots, d$, and it is assumed that

$$\|q\|_{0,\infty,M} \leq C\,h_M^{-\frac{d}{2}}\,\|q\|_{0,M} \qquad \forall\ q \in G_M,\ M \in \mathscr{M}_h\,. \tag{5}$$

To characterize the approximation properties of the spaces W_h and D_M, it is assumed that there exist interpolation operators $i_h \in \mathscr{L}(H^2(\Omega), W_h) \cap \mathscr{L}(H^2(\Omega) \cap H_0^1(\Omega), V_h)$ and $j_M \in \mathscr{L}(H^1(M), D_M)$, $M \in \mathscr{M}_h$, such that, for some constants $l \in \mathbb{N}$ and $C > 0$ and for any set $M \in \mathscr{M}_h$, it holds

$$|v - i_h v|_{1,M} + h_M^{-1}\,\|v - i_h v\|_{0,M} \leq C\,h_M^k\,|v|_{k+1,M} \quad \forall\ v \in H^{k+1}(M),\ k = 1, \ldots, l\,, \tag{6}$$

$$\|q - j_M q\|_{0,M} \leq C\,h_M^k\,|q|_{k,M} \quad \forall\ q \in H^k(M),\ k = 1, \ldots, l\,. \tag{7}$$

We refer to [3] for examples of spaces W_h and D_M possessing the properties formulated in this section.

3 A Local Projection Discretization

The weak form of problem (1) is: Find $u \in H^1(\Omega)$ such that $u = u_b$ on $\partial\Omega$ and

$$a(u, v) = (f, v) \qquad \forall\ v \in H_0^1(\Omega)\,, \tag{8}$$

where $(\cdot,\cdot)$ denotes the inner product in $L^2(\Omega)$ and the bilinear form a is given by

$$a(u, v) = \varepsilon\,(\nabla u, \nabla v) + (\mathbf{b}\cdot\nabla u, v) + (c\,u, v)\,.$$

For any $M \in \mathscr{M}_h$, a continuous linear projection operator π_M is introduced which maps the space $L^2(M)$ onto the space D_M. It is assumed that $\|\pi_M\|_{\mathscr{L}(L^2(M),L^2(M))} \le C$ for any $M \in \mathscr{M}_h$. Using this operator, the fluctuation operator $\kappa_M := id - \pi_M$ is defined, where id is the identity operator on $L^2(M)$. Then, clearly

$$\|\kappa_M\|_{\mathscr{L}(L^2(M),L^2(M))} \le C \qquad \forall\ M \in \mathscr{M}_h\,. \tag{9}$$

Since κ_M vanishes on D_M, it follows from (9) and (7) that

$$\|\kappa_M\, q\|_{0,M} \le C\, h_M^k\, |q|_{k,M} \qquad \forall\ q \in H^k(M),\ M \in \mathscr{M}_h,\ k = 0, \ldots, l\,. \tag{10}$$

An application of κ_M to a vector-valued function means that κ_M is applied componentwise.

For any $M \in \mathscr{M}_h$, a constant $\mathbf{b}_M \in \mathbb{R}^d$ is chosen such that

$$|\mathbf{b}_M| \le \|\mathbf{b}\|_{0,\infty,M}\,, \qquad \|\mathbf{b} - \mathbf{b}_M\|_{0,\infty,M} \le C\, h_M\, |\mathbf{b}|_{1,\infty,M}\,. \tag{11}$$

A typical choice for $\mathbf{b}_M$ is the value of $\mathbf{b}$ at one point of M, or the integral mean value of $\mathbf{b}$ over M. In addition, a function $\widetilde{u}_{bh} \in W_h$ is introduced such that its trace approximates the boundary condition u_b.

We are now ready to present the finite element method to be studied: Find $u_h \in W_h$ such that $u_h - \widetilde{u}_{bh} \in V_h$ and

$$a(u_h, v_h) + s_h(u_h, v_h) + d_h(u_h; u_h, v_h) = (f, v_h) \qquad \forall\ v_h \in V_h\,, \tag{12}$$

where

$$s_h(u, v) = \sum_{M\in\mathscr{M}_h} \tau_M\,(\kappa_M(\mathbf{b}_M\cdot\nabla u), \kappa_M(\mathbf{b}_M\cdot\nabla v))_M\,,$$

$$d_h(w; u, v) = \sum_{M\in\mathscr{M}_h} (\tau_M^{\mathrm{sold}}(w)\,\kappa_M(P_M\nabla u), \kappa_M(P_M\nabla v))_M\,,$$

$(\cdot,\cdot)_M$ is the inner product in $L^2(M)$ and $P_M : \mathbb{R}^d \to \mathbb{R}^d$ is the orthogonal projection onto the line (plane) orthogonal to $\mathbf{b}_M$. The stabilization parameters are given by

$$\tau_M = \tau_0 \min\left\{ \frac{h_M}{\|\mathbf{b}\|_{0,\infty,M}}, \frac{h_M^2}{\varepsilon} \right\},$$

$$\tau_M^{\text{sold}}(u_h) = \begin{cases} \beta\, h_M\, |\mathbf{b}_M|\, \dfrac{h_M^d\, |\kappa_M(P_M \nabla u_h)|^2}{|u_h|_{1,M}^2} & \text{if } |u_h|_{1,M} \neq 0, \\ 0 & \text{if } |u_h|_{1,M} = 0, \end{cases}$$

where τ_0 and β are positive constants.

Remark 1. Using (11), (9), and $\|P_M\|_2 = 1$, one obtains

$$\|\tau_M^{\text{sold}}(v)\|_{0,1,M} \leq C\, h_M^{1+d}\, \|\mathbf{b}\|_{0,\infty,M} \qquad \forall\, v \in H^1(\Omega),\ M \in \mathscr{M}_h\,. \tag{13}$$

In the analysis, the error will be measured using the following mesh-dependent norm

$$\|v\|_{\text{LPS}} := \left(\varepsilon\, |v|_{1,\Omega}^2 + \|\sigma^{1/2}\, v\|_{0,\Omega}^2 + s_h(v,v) \right)^{1/2}.$$

Note that integrating by parts gives

$$a(v,v) + s_h(v,v) = \|v\|_{\text{LPS}}^2 \qquad \forall\, v \in H_0^1(\Omega)\,. \tag{14}$$

4 Well-Posedness of the Nonlinear Discrete Problem

This section studies the existence of solutions for the nonlinear discrete problem (12). Let us define the nonlinear operator $T_h : V_h \to V_h$ by

$$(T_h z_h, v_h) = a(z_h + \widetilde{u}_{bh}, v_h) + s_h(z_h + \widetilde{u}_{bh}, v_h) + d_h(z_h + \widetilde{u}_{bh}; z_h + \widetilde{u}_{bh}, v_h) - (f, v_h)$$

for any $z_h, v_h \in V_h$. Then $u_h \in W_h$ is a solution of (12) if and only if $u_h|_{\partial\Omega} = \widetilde{u}_{bh}|_{\partial\Omega}$ and $T_h(u_h - \widetilde{u}_{bh}) = 0$. Thus, our aim is to prove that the operator T_h has a zero in V_h. To this end, we shall use the following simple consequence of Brouwer's fixed-point theorem.

Lemma 1. *Let X be a finite-dimensional Hilbert space with inner product $(\cdot,\cdot)$ and norm $\|\cdot\|$. Let $P : X \to X$ be a continuous mapping and $K > 0$ a real number such that $(Px, x) > 0$ for any $x \in X$ with $\|x\| = K$. Then there exists $x \in X$ such that $\|x\| \leq K$ and $Px = 0$.*

Proof. See [6, p. 164, Lemma 1.4].

Theorem 1. *The problem (12) has a solution.*

Proof. In view of (14), for any $z_h \in V_h$, it holds

$$(T_h z_h, z_h) = \|z_h\|^2_{\mathrm{LPS}} + d_h(z_h + \widetilde{u}_{bh}; z_h, z_h) + a(\widetilde{u}_{bh}, z_h) + s_h(\widetilde{u}_{bh}, z_h) + d_h(z_h + \widetilde{u}_{bh}; \widetilde{u}_{bh}, z_h) - (f, z_h).$$

According to (13), one has

$$|d_h(u; v, z)| \le C \sum_{M \in \mathscr{M}_h} h_M^{1+d} \, \|\mathbf{b}\|_{0,\infty,M} \, \|\kappa_M (P_M \nabla v)\|_{0,\infty,M} \, \|\kappa_M (P_M \nabla z)\|_{0,\infty,M}$$

for any $u, v, z \in W^{1,\infty}(\Omega)$. Thus, applying (5), (9), the equivalence of norms on finite-dimensional spaces, the Cauchy-Schwarz inequality, and the Young inequality, one deduces that

$$(T_h z_h, z_h) \ge \frac{1}{2} \|z_h\|^2_{\mathrm{LPS}} - C_0 \left(\|\widetilde{u}_{bh}\|^2_{0,\Omega} + \|f\|^2_{0,\Omega}\right),$$

where $C_0 > 0$ depends on ε, $\mathbf{b}$, c, σ_0, h, and W_h but not on z_h. Consequently,

$$(T_h z_h, z_h) \ge C_1 \|z_h\|^2_{0,\Omega} - C_2 \qquad \forall\, z_h \in V_h\,,$$

where C_1, C_2 are positive constants. Thus, in view of Lemma 1 with any $K > \sqrt{C_2/C_1}$, the operator T_h has a zero and hence the problem (12) has a solution.

5 Error Estimate

Lemma 2. *There exists an operator $\varrho_h : L^2(\Omega) \to V_h$ such that, for any $v, w \in L^2(\Omega)$, the following estimates hold*

$$|(v - \varrho_h v, w)| \le C \sum_{M \in \mathscr{M}_h} \|v\|_{0,M} \, \|\kappa_M w\|_{0,M}\,, \tag{15}$$

$$|\varrho_h v|^2_{1,M} + h_M^{-2} \|\varrho_h v\|^2_{0,M} \le C \sum_{\substack{M' \in \mathscr{M}_h, \\ M \cap M' \neq \emptyset}} h_{M'}^{-2} \|v\|^2_{0,M'} \qquad \forall\, M \in \mathscr{M}_h\,. \tag{16}$$

Proof. See [4, Lemma 1].

Using the operators i_h and ϱ_h, we introduce the operator $r_h \in \mathscr{L}(H^2(\Omega), W_h) \cap \mathscr{L}(H^2(\Omega) \cap H^1_0(\Omega), V_h)$ by $r_h v := i_h v + \varrho_h(v - i_h v)$. To formulate the interpolation properties of r_h, it is convenient to introduce the mesh dependent norm

$$\|v\|_{1,h} = \left(\sum_{M \in \mathscr{M}_h} \{ |v|^2_{1,M} + h_M^{-2} \, \|v\|^2_{0,M} \} \right)^{1/2} .$$

Then, using (16), (2), (3), and (6), one obtains

$$\|v - r_h v\|_{1,h} \le C \, \|v - i_h v\|_{1,h} \le \tilde{C} \, h^k \, |v|_{k+1,\Omega} \qquad \forall \; v \in H^{k+1}(\Omega), \; k = 1, \ldots, l \, . \tag{17}$$

Lemma 3. *Let $u \in H^{k+1}(\Omega)$ for some $k \in \{1, \ldots, l\}$, and let $\eta := u - r_h u$. Then, for any $v_h \in V_h \setminus \{0\}$, the following estimate holds*

$$\|\eta\|_{\mathrm{LPS}} + \frac{a(\eta, v_h) + s_h(\eta, v_h) - s_h(u, v_h)}{\|v_h\|_{\mathrm{LPS}}} \le C \left(\varepsilon + h \, \|\mathbf{b}\|_{0,\infty,\Omega} + h^2 \, \|\sigma\|_{0,\infty,\Omega} + h^2 \, |\mathbf{b}|^2_{1,\infty,\Omega} \, \sigma_0^{-1} \right)^{1/2} h^k \, |u|_{k+1,\Omega} \, . \tag{18}$$

Proof. See [4].

Lemma 4. *For any $w_h \in W_h$ and $u \in H^{k+1}(\Omega)$ with $k \in \{1, \ldots, l\}$, it holds*

$$d_h(w_h; r_h u, r_h u) \le C \, \|\mathbf{b}\|_{0,\infty,\Omega} \, h^{2k+1} \, |u|^2_{k+1,\Omega} \, . \tag{19}$$

Proof. First, the application of (5), (13), and (3) leads to

$$\begin{aligned} d_h(w_h; r_h u, r_h u) &\le \sum_{M \in \mathscr{M}_h} \|\tau_M^{\mathrm{sold}}(w_h)\|_{0,1,M} \, \|\kappa_M (P_M \nabla (r_h u))\|^2_{0,\infty,M} \\ &\le C \, h \, \|\mathbf{b}\|_{0,\infty,\Omega} \sum_{M \in \mathscr{M}_h} \|\kappa_M \nabla (r_h u)\|^2_{0,M} \, . \end{aligned}$$

Using (9) and (10), for $u \in H^{k+1}(\Omega)$ with $k \in \{1, \ldots, l\}$ there holds

$$\begin{aligned} \|\kappa_M \nabla (r_h u)\|_{0,M} &\le \|\kappa_M \nabla u\|_{0,M} + \|\kappa_M \nabla (u - r_h u)\|_{0,M} \\ &\le C \, h_M^k \, |u|_{k+1,M} + C \, |u - r_h u|_{1,M} \, . \end{aligned}$$

Thus, (19) follows from (2), (3), and (17).

We are now in a position to prove the main result of this paper.

Theorem 2. *Let the weak solution of (1) satisfy $u \in H^{k+1}(\Omega)$ for some $k \in \{1, \ldots, l\}$. Let $\widetilde{u}_b \in H^2(\Omega)$ be an extension of u_b and let $\widetilde{u}_{bh} = i_h \widetilde{u}_b$. Then the solution u_h of the local projection discretization (12) satisfies the error estimate*

$$\|u - u_h\|_{\mathrm{LPS}} \le C \left(\varepsilon + h \, \|\mathbf{b}\|_{0,\infty,\Omega} + h^2 \, \|\sigma\|_{0,\infty,\Omega} + h^2 \, |\mathbf{b}|^2_{1,\infty,\Omega} \, \sigma_0^{-1} \right)^{1/2} h^k \, |u|_{k+1,\Omega} \, .$$

Proof. Set $\eta := u - r_h u$ and $e_h := u_h - r_h u$. From (12) and (8), it follows that

$$\begin{aligned} a(e_h, e_h) + s_h(e_h, e_h) + d_h(u_h; u_h, e_h) \\ &= a(u_h, e_h) + s_h(u_h, e_h) + d_h(u_h; u_h, e_h) - a(r_h u, e_h) - s_h(r_h u, e_h) \\ &= a(\eta, e_h) + s_h(\eta, e_h) - s_h(u, e_h)\,. \end{aligned}$$

Thus, in view of (14), one gets

$$\|e_h\|_{\mathrm{LPS}}^2 + d_h(u_h; e_h, e_h) = a(\eta, e_h) + s_h(\eta, e_h) - s_h(u, e_h) - d_h(u_h; r_h u, e_h)\,.$$

The first three terms on the right-hand side can be estimated using (18). Applying Hölder's and Young's inequalities, one gets $d_h(u_h; r_h u, e_h) \leq d_h(u_h; r_h u, r_h u) + \frac{1}{4} d_h(u_h; e_h, e_h)$. Therefore, using (19), one obtains

$$\begin{aligned} \|e_h\|_{\mathrm{LPS}}^2 + d_h(u_h; e_h, e_h) \\ &\leq C \left(\varepsilon + h \|\mathbf{b}\|_{0,\infty,\Omega} + h^2 \|\sigma\|_{0,\infty,\Omega} + h^2 |\mathbf{b}|_{1,\infty,\Omega}^2 \sigma_0^{-1} \right) h^{2k} |u|_{k+1,\Omega}^2\,. \end{aligned}$$

Finally, using the triangle inequality and the estimate (18), the statement of the theorem follows.

6 Numerical Results

In this section we illustrate the properties of the method proposed in this paper by numerical results obtained for the following example.

Example 1. Solution with two interior layers. Equation (1) is considered with $\Omega = (0,1)^2$, $\varepsilon = 10^{-8}$, $\mathbf{b}(x, y) = (-y, x)^T$, $c = f = 0$, and the boundary conditions

$$u = u_b \quad \text{on } \Gamma^D\,, \qquad \frac{\partial u}{\partial \mathbf{n}} = 0 \quad \text{on } \Gamma^N\,,$$

where $\Gamma^N = \{0\} \times (0,1)$, $\Gamma^D = \partial\Omega \setminus \overline{\Gamma^N}$, $\mathbf{n}$ is the outward pointing unit normal vector to the boundary of Ω, and

$$u_b(x, y) = \begin{cases} 1 & \text{for } (x, y) \in (1/3, 2/3) \times \{0\}, \\ 0 & \text{else on } \Gamma^D. \end{cases}$$

We used a triangulation $\mathscr{T}_h$ of Ω constructed by dividing Ω into 32×32 equal squares and each square into two triangles by drawing a diagonal from bottom left to top right. Each set $M \in \mathscr{M}_h$ is the union of all triangles of $\mathscr{T}_h$ possessing a common

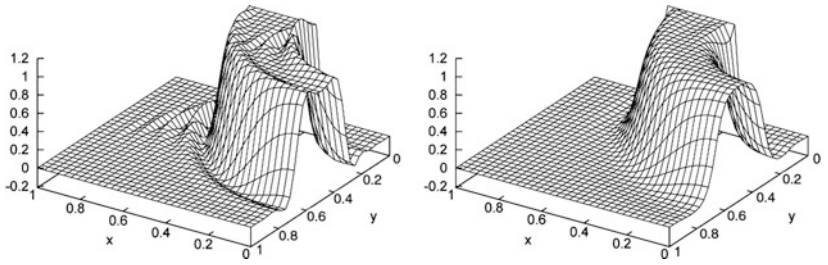


Fig. 1 LPS solutions for $\tau_0 = 0.02$, $\beta = 0$ (*left*) and $\tau_0 = 0.02$, $\beta = 0.05$ (*right*)

interior vertex of $\mathscr{T}_h$. Thus the sets from $\mathscr{M}_h$ generally overlap. The space W_h consists of continuous piecewise linear functions and the spaces D_M are spaces of constant functions. Figure 1 shows that the crosswind diffusion term d_h leads to a reduction of spurious oscillations compared to the standard linear LPS method.

Acknowledgements The work of P. Knobloch is a part of the research project MSM 0021620839 financed by MSMT and it was partly supported by the Grant Agency of the Czech Republic under the grant No. 201/08/0012.

References

1. S. Ganesan and L. Tobiska. Stabilization by local projection for convection-diffusion and incompressible flow problems. *J. Sci. Comput.*, 43:326–342, 2010.
2. V. John and P. Knobloch. On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations: Part I – A review. *Comput. Methods Appl. Mech. Engrg.*, 196:2197–2215, 2007.
3. P. Knobloch. A generalization of the local projection stabilization for convection–diffusion–reaction equations. *SIAM J. Numer. Anal.*, 48:659–680, 2010.
4. P. Knobloch. Local projection method for convection–diffusion–reaction problems with projection spaces defined on overlapping sets. In G. Kreiss, P. Lötstedt, A. Målqvist, and M. Neytcheva, editors, *Numerical Mathematics and Advanced Applications 2009*, pages 497–505. Springer-Verlag, Berlin, 2010.
5. H.-G. Roos, M. Stynes, and L. Tobiska. *Robust Numerical Methods for Singularly Perturbed Differential Equations. Convection–Diffusion–Reaction and Flow Problems. 2nd ed.* Springer-Verlag, Berlin, 2008.
6. R. Temam. *Navier–Stokes Equations. Theory and numerical analysis*. North–Holland, Amsterdam, 1977.

An Improved Optimal Order Mixed Finite Element Method for Semilinear Transport Problems

M. Bause, F. Brunner, F.A. Radu, and P. Knabner

Abstract We propose and study the numerical approximation of an advection-diffusion-reaction model equation by a modified Brezzi–Douglas–Marini mixed finite element method. Nonlinear advection is admitted, arising in complex and coupled flow and transport systems. In contrast to the classical variant of this approach, optimal second-order convergence of the scalar and the vector variable is ensured. No loss of rate of convergence due to the presence of the advection term is observed.

1 Introduction

A mathematical prototye model for describing reactive solute transport in porous media is given by the advection-diffusion-reaction equation

$$\partial_t(\Theta c) - \nabla \cdot (\boldsymbol{D}\nabla c - \boldsymbol{Q}c) = \Theta r(c) \tag{1}$$

with c denoting the concentration of the solute, $\boldsymbol{D}$ the diffusion/dispersion matrix, $r(\cdot)$ a reaction rate and Θ and $\boldsymbol{Q}$ the water saturation and water flux, respectively. The accurate and reliable numerical approximation of such models is an active area

M. Bause
University of the Federal Armed Forces, Holstenhofweg 85, D-22043, Hamburg, Germany
e-mail: bause@hsu-hh.de

F. Brunner · P. Knabner
Department of Mathematics, University of Erlangen-Nuremberg, Cauerstr. 11, D-91058, Erlangen, Germany
e-mail: brunner@am.uni-erlangen.de; knabner@am.uni-erlangen.de

F.A. Radu (✉)
Department of Mathematics, University of Bergen, P. O. Box 7800, N-5020, Bergen, Norway
e-mail: Florin.Radu@math.uib.no

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_27,

of research because of its potential for practical applications. Such models can be found in environmental sciences and civil engineering as well as in various industrial processes involving active filtration. For instance, groundwater contamination from leaky underground storage tanks, chemical pills and various human activities is one serious problem today. Here, Eq. (1) can be used to study the spread of pollution released in the subsurface and assess the danger.

Regarding the numerical approximation of solutions of Eq. (1), mixed finite element approaches were applied successfully in the past. They provide locally mass conservative quantities, a property critical for the transport problem (1) in order to avoid creating artificial mass sources and sinks. Moreover, higher order discretizations of the unknowns offer advantages over their lowest-order counterparts. Higher order methods have shown to introduce a smaller amount of numerical diffusion, a property essential for reactive multicomponent transport processes with mixing of chemical species; cf. [2, 4, 8]. The lowest-order Brezzi–Douglas–Marini mixed finite element method (BDM_1) is the algorithmically simplest mixed approach that is *formally* capable to provide second-order accurate approximations of the scalar unknown and the vectorial (flux) unknown; cf. [3]. However, it is well known (cf. [5]) that the BDM_1 method applied to an advection-diffusion equation in mass conservative form, i.e. the flux variable is defined by the total flux

$$\boldsymbol{q} = -\boldsymbol{D}\nabla c + \boldsymbol{Q}c\,, \tag{2}$$

is suboptimal of first-order accuracy only with respect to the approximation of the flux variable. In applications of practical interest, mathematical models involving interfaces or transmissions of unknowns on interior boundaries are also studied (cf. [6, 12]) or iterative subdomain methods (e.g. domain decomposition methods) are applied. In these cases optimal higher order approximations of the flux variable are desirable which requires modifications to the standard mixed approaches. In [4] we proposed a modification of the classical BDM_1 scheme; cf. Sect. 2. Optimal order of convergence of the scalar and vectorial variable was observed numerically for a wide class of problems of practical interest. However, a rigorous error analysis for this scheme is still an open problem. The idea used in [4] comes from works on the lowest-order Raviart–Thomas mixed finite element method; cf. [9, 10, 14]. It is based on the fact that the Lagrange multipliers of a hybrid mixed formulation can be used for constructing a second-order accurate approximation of the scalar variable (concentration); cf. [3]. In [4], we used the Lagrange multipliers to discretize the advective part of the total flux and, thereby, we improved the order of convergence of the BDM_1 method. This modification ensures that the advective part of the discrete flux belongs to the function space $\boldsymbol{H}(div, \Omega)$, which is not the case for the classical BDM_1 scheme or for the lowest-order Raviart–Thomas method.

However, in all the numerical experiments presented in [4] it was tacitly assumed that the prescribed flow field does not depend on the concentration variable. This means that the water saturation Θ and the water flux $\boldsymbol{Q}$ were prescribed independently from the species concentration c or that the model equations governing the

underlying flow field decouple from the transport problem. In many applications of practical interest, e.g. surfactant transport in porous media [7, 11, 13], this simplification is not sufficiently satisfied and the flow field depends on the dissolved species concentration. Therefore and due to the great importance of the approximation of the advective model term with respect to the numerical performance and convergence properties of the overall mixed approximation scheme, we found it worthwhile and necessary to study the modified BDM_1 approach also for nonlinear advective terms. Precisely, we assume that

$$\Theta = \Theta(\boldsymbol{x}, c), \qquad \boldsymbol{Q} = \boldsymbol{Q}(\boldsymbol{x}, c) \tag{3}$$

is satisfied. By (1), (3) we mimic coupled flow and transport systems. The numerical results that are presented in Sect. 3 for a single semilinear transport equation and for coupled systems of such equations with nonlinear advection nicely confirm our conjecture that the modified BDM_1 scheme (cf. Sect. 2) yields optimal second-order convergence for the scalar and the vectorial variable and, thereby, is superior to the classical BDM_1 approach.

2 The Modified BDM_1 Scheme

In this section we briefly sketch the discretization of Eq. (1) by the classical and the modified BDM_1 mixed finite element method in space and the backward Euler scheme in time. An extension to a full multicomponent system is straightforward. Rewriting Eq. (1) in its mixed form reads as

$$\partial_t(\Theta c) + \nabla \cdot \boldsymbol{q} = \Theta r, \quad \boldsymbol{q} = -D\nabla c + \boldsymbol{Q}c \quad \text{in } J \times \Omega.$$

Here, the domain $\Omega \subset \mathbb{R}^2$ is assumed to be polygonally bounded and $J = (0, T]$ is a finite time interval. The equation is supplemented with homogeneous Dirichlet boundary conditions and the initial condition $c(0, \cdot) = c_0(\cdot)$ for a given function c_0. Now, let $\mathscr{T}_h$ be a decomposition of Ω into closed triangles $K \in \mathscr{T}_h$ and let $\mathscr{E}$ denote the set of all edges and $\mathscr{E}_D$ the set of edges on the Dirichlet part of the boundary. The discretization in time is done by the backward Euler method, where the time interval J is divided in N subintervals $(t_{n-1}, t_n]$, $n \in \{1, \ldots, N\}$, with $\tau_n = t_n - t_{n-1}$ denoting the time step size. For the spatial discretization with the BDM_1 mixed finite element method we define the spaces

$$\boldsymbol{V}_h = \{\boldsymbol{q} \in \boldsymbol{H}(\mathrm{div}; \Omega) \mid \boldsymbol{q}|_K \in BDM_1(K)\ \forall K \in \mathscr{T}_h\},$$
$$W_h = \{w \in L^2(\Omega) \mid w|_K \in \mathscr{P}_0(K)\ \forall K \in \mathscr{T}_h\},$$

where $BDM_1(K) = (\mathscr{P}_1(K))^2$ and $\mathscr{P}_k(K)$ is the space of polynomials of degree at most k on K. Finally, we can define the fully discrete mixed weak formulation of the reactive transport equation at time level n:

Problem 1. Let c_h^{n-1} be given. Find $(\boldsymbol{q}_h^n,\ c_h^n) \in \boldsymbol{V}_h \times W_h$ satisfying

$$\langle \boldsymbol{D}^{-1}\boldsymbol{q}_h^n, \boldsymbol{v}_h\rangle_\Omega - \langle c_h^n, \nabla\cdot\boldsymbol{v}_h\rangle_\Omega - \langle \boldsymbol{D}^{-1}\boldsymbol{Q}^n c_h^n, \boldsymbol{v}_h\rangle_\Omega = 0,$$

$$\frac{1}{\tau^n}\langle \Theta^n c_h^n - \Theta^{n-1}c_h^{n-1}, w_h\rangle_\Omega + \langle \nabla\cdot\boldsymbol{q}_h^n, w_h\rangle_\Omega = \langle \Theta^n r^n, w_h\rangle_\Omega$$

for all $\boldsymbol{v}_h \in \boldsymbol{V}_h$ and $w_h \in W_h$.

The superscript n refers to the evaluation of a variable at time $t = t_n$. To improve the properties of the linear systems that remain to be solved after a linearization of the discrete equations and to reduce the number of global unknowns, a hybridization technique is used; cf. [3, 9]. The condition $\boldsymbol{V}_h \subseteq \boldsymbol{V}$ is relaxed and the space $\boldsymbol{V}_h$ is replaced by the augmented space $\tilde{\boldsymbol{V}}_h = \{\boldsymbol{q} \in \boldsymbol{L}^2(\Omega) \mid \boldsymbol{q}|_K \in BDM_1(K)\ \forall K \in \mathscr{T}_h\}$. The continuity of the normal fluxes over interelement edges, which is no longer incorporated in the function space, is ensured by introducing Lagrange multipliers from the space

$$\Lambda_h = \{\lambda \in L^2(\mathscr{E}) \mid \lambda|_E \in \mathscr{P}_1(E)\ \forall E \in \mathscr{E}\}$$

and requiring an additional variational equation. The resulting mixed hybrid discrete formulation is equivalent to the non-hybrid form and reads as:

Problem 2. Let c_h^{n-1} be given. Find $(\boldsymbol{q}_h^n,\ c_h^n,\ \lambda_h^n) \in \tilde{\boldsymbol{V}}_h \times W_h \times \Lambda_h$ satisfying

$$\langle \boldsymbol{D}^{-1}\boldsymbol{q}_h^n, \boldsymbol{v}_h\rangle_\Omega - \langle c_h^n, \nabla\cdot\boldsymbol{v}_h\rangle_\Omega - \langle \boldsymbol{D}^{-1}\boldsymbol{Q}^n c_h^n, \boldsymbol{v}_h\rangle_\Omega = -\sum_{K\in\mathscr{T}_h}\langle \lambda_h^n, \boldsymbol{v}_h\cdot\boldsymbol{n}\rangle_{\partial K}$$

$$\frac{1}{\tau^n}\langle \Theta^n c_h^n - \Theta^{n-1}c_h^{n-1}, w_h\rangle_\Omega + \langle \nabla\cdot\boldsymbol{q}_h^n, w_h\rangle_\Omega = \langle \Theta^n r^n, w_h\rangle_\Omega,$$

$$\sum_{K\in\mathscr{T}_h}\langle \mu_h, \boldsymbol{q}_h^n\cdot\boldsymbol{n}\rangle_{\partial K} = 0$$

for all $\boldsymbol{v}_h \in \tilde{\boldsymbol{V}}_h$, $w_h \in W_h$ and $\mu_h \in \Lambda_h$,

where $\boldsymbol{n}$ denotes the outer unit normal to ∂K. For an algebraic formulation of Problem 2, basis functions of the involved function spaces are needed. There exist basis functions $\{\boldsymbol{v}_{KE}^{(i)}\}_{K\in\mathscr{T}_h, E\in\mathscr{E}, i=1,2}$ of $\tilde{\boldsymbol{V}}_h$ and $\{\mu_E^{(j)}\}_{E\in\mathscr{E},\ j=1,2}$ of Λ_h such that

$$\langle \nabla\cdot\boldsymbol{v}_{KE}^{(i)}, 1\rangle_K = 1, \quad \langle \boldsymbol{v}_{KE'}^{(i)}\cdot\boldsymbol{n}_E, 1\rangle_E = \delta_{EE'}, \quad \langle \mu_E^{(j)}\,\boldsymbol{v}_{KE'}^{(i)}\cdot\boldsymbol{n}_E, 1\rangle_E = \delta_{EE'}\delta_{ij}$$

for $E, E' \subset \partial K,\ i, j = 1, 2$. The basis functions for the scalar variables are given by respective characteristic functions χ_K on the triangle K. Finally, the unknowns are written in terms of the basis functions,

$$\boldsymbol{q}_h^n = \sum_{K\in\mathscr{T}_h}\sum_{E\subset\partial K}\sum_{j=1}^{2} q_{KE_j}^n \boldsymbol{v}_{KE}^{(j)}, \quad \lambda_h^n = \sum_{E\in\mathscr{E}}\sum_{j=1}^{2}\lambda_{E_j}^n \mu_E^{(j)}, \quad c_h^n = \sum_{K\in\mathscr{T}_h} c_K^n \chi_K,$$

and the basis functions are applied as test functions in the discrete mixed formulation to obtain the following algebraic system of nonlinear discrete equations:

Equations for the fluxes:

$$\sum_{E\subset\partial K}\sum_{j=1}^{2} q_{KE_j}^n \langle \boldsymbol{D}^{-1}\boldsymbol{v}_{KE}^{(j)}, \boldsymbol{v}_{KE'}^{(l)}\rangle_K - c_K^n - \sum_{E\subset\partial K}\sum_{j=1}^{2} Q_{KE_j}^n c_K^n \langle \boldsymbol{D}^{-1}\boldsymbol{v}_{KE}^{(j)}, \boldsymbol{v}_{KE'}^{(l)}\rangle_K = -\lambda_{E_l'}^n \quad (4)$$

$$\forall K\in\mathscr{T}_h,\ E'\subset\partial K,\ l=1,2.$$

Mass conservation equations:

$$\frac{1}{\tau^n}(\Theta_K^n c_K^n - \Theta_K^{n-1} c_K^{n-1}) + \sum_{E\subset\partial K}\sum_{j=1}^{2} q_{KE_j}^n = r_K^n \Theta_K^n \quad \forall K\in\mathscr{T}_h, \quad (5)$$

where $\Theta_K^n := \int_K \Theta^n\, dx$, $r_K^n := r(c_K^n)$.

Equations for the Lagrange multipliers:

$$\sum_{\substack{K\in\mathscr{T}_h:\\ E\subset\partial K}} q_{KE_j}^n = 0 \quad \forall E\in\mathscr{E}\setminus\mathscr{E}_D,\ j=1,2. \quad (6)$$

Here, we assume that the water flux can be written as $\boldsymbol{Q}(\boldsymbol{x},c) = \hat{\boldsymbol{Q}}(\boldsymbol{x})f(c)$ with $\hat{\boldsymbol{Q}}\in\boldsymbol{V}_h$ and a function $f:\mathbb{R}\to\mathbb{R}$ modeling the feedback of the concentrations on the water flux. The coefficients $Q_{KE_j}^n$ in (4) are then defined as $Q_{KE_j}^n := \hat{Q}_{KE_j} f_K^n$ with $f_K^n := f(c_K^n)$, where $\hat{Q}_{KE_j}$ denote the coefficients of the representation of $\hat{\boldsymbol{Q}}$ in the local BDM_1 basis. Defining the matrices $(\boldsymbol{B}_K)_{E_iE_j'} := \langle \boldsymbol{D}^{-1}\boldsymbol{v}_{KE}^{(i)}, \boldsymbol{v}_{KE'}^{(j)}\rangle_K$ and inserting them into the equations for the fluxes (4) yields that

$$\sum_{E\subset\partial K}\sum_{j=1}^{2} q_{KE_j}^n \boldsymbol{B}_{KE_jE_l'} = c_K^n - \lambda_{E_l'}^n + \sum_{E\subset\partial K}\sum_{j=1}^{2}\boldsymbol{B}_{KE_jE_l'}\hat{Q}_{KE_j} f_K^n c_K^n \quad (7)$$

$$\forall K\in\mathscr{T}_h,\ E'\subset\partial K,\ l=1,2.$$

This set of equations represents a linear system of equations on each element from which the flux variables can be eliminated explicitly by

$$q^n_{KE_j} = \sum_{E'\subset\partial K}\sum_{l=1}^{2}(\boldsymbol{B}_K^{-1})_{E'_l E_j}(c^n_K - \lambda^n_{E'_l}) + \hat{Q}_{KE_j} f^n_K c^n_K, \tag{8}$$

where $\boldsymbol{B}_K^{-1}$ is the inverse of the matrix $\boldsymbol{B}_K = (\boldsymbol{B}_K)_{E_i E'_j}$. Using this relation in Eq. (5), the mass conservation equations read as

$$\frac{1}{\tau^n}(\Theta^n_K c^n_K - \Theta^{n-1}_K c^{n-1}_K) + \sum_{E,E'\subset\partial K}\sum_{j,l=1}^{2}(\boldsymbol{B}_K^{-1})_{E'_l E_j}(c^n_K - \lambda^n_{E'_l})$$

$$+ c^n_K f^n_K \sum_{E\subset\partial K}\sum_{j=1}^{2}\hat{Q}_{KE_j} = \Theta^n_K r^n_K \quad \forall K \in \mathscr{T}_h. \tag{9}$$

These nonlinear equations are solved on each element by a nested Newton iteration to determine the values $c^n_K, K \in \mathscr{T}_h$. Finally, the concentrations and fluxes are inserted into (6) such that the following overall global system of equations for the Lagrange multipliers is obtained:

$$\sum_{\substack{K\in\mathscr{T}_h:\\ E\subset\partial K}}\left\{\sum_{E'\subset\partial K}\sum_{l=1}^{2}(\boldsymbol{B}_K^{-1})_{E'_l E_j}(c^n_K - \lambda^n_{E'_l}) + \hat{Q}_{KE_j} f^n_K c^n_K\right\} = 0 \quad \forall E \in \mathscr{E}\setminus\mathscr{E}_D,\ j = 1,2.$$

Once the global system for the Lagrange multipliers has been solved, the flux unknowns can be reconstructed efficiently on each element from Eq. (8).

In [5], it was shown that the classical BDM_1 scheme yields only a first-order accurate approximation of the total flux $\boldsymbol{q} = -\boldsymbol{D}\nabla c + \boldsymbol{Q}c$ in the L^2 norm in the case of presence of an advection term $\boldsymbol{Q}c$. A modification of the discrete equations using the Lagrange multipliers for the discretization of the flux provided optimal second-order convergence in the numerical experiments reported in [4]. It is our goal here to analyze if this improvement in the convergence property is preserved when there is a nonlinear feedback of the concentration on the water flux and the saturation.The **equations for the fluxes in the modified scheme** now read as:

$$\sum_{E\subset\partial K}\sum_{j=1}^{2} q^n_{KE_j}\boldsymbol{B}_{KE_jE'_l} = c^n_K - \lambda^n_{E'_l} + \sum_{E\subset\partial K}\sum_{j=1}^{2}\boldsymbol{B}_{KE_jE'_l}\hat{Q}^n_{KE_j} f(\lambda^n_h(\boldsymbol{x}^j_E))\,\lambda^n_h(\boldsymbol{x}^j_E)$$

$$\forall K \in \mathscr{T}_h,\ E' \subset \partial K,\ l = 1,2.$$

Here, by $\boldsymbol{x}^i_E$ we denote the nodes of the triangle K corresponding to the degrees of freedom of the flux basis functions.

Table 1 Calculated errors and experimental convergence rates for the mixed advection-diffusion equation using the modified BDM_1 scheme

k	$\|c - c_h\|$	α_k	$\|c - \tilde{c}_h\|$	α_k	$\|\boldsymbol{q} - \boldsymbol{q}_h\|$	α_h
0	3.027e − 02		4.012e − 02		7.381e − 02	
1	1.684e − 02	0.85	1.215e − 02	1.72	3.874e − 02	0.93
2	8.698e − 03	0.95	3.513e − 03	1.79	1.201e − 02	1.69
3	4.382e − 03	0.99	9.172e − 04	1.94	3.219e − 03	1.90
4	2.195e − 03	1.00	2.320e − 04	1.98	8.259e − 04	1.96
5	1.098e − 03	1.00	5.816e − 05	2.00	2.087e − 04	1.98
6	5.490e − 04	1.00	1.455e − 05	2.00	5.243e − 05	1.99
7	2.745e − 04	1.00	3.639e − 06	2.00	1.314e − 05	2.00

3 Numerical Studies

3.1 Stationary Advection-Diffusion Equation

The simplest application for which the modified scheme becomes relevant is the mixed advection-diffusion problem

$$\nabla \cdot \boldsymbol{q} = F, \quad \boldsymbol{q} = -D\nabla c + \boldsymbol{Q}c \quad \text{in } \Omega.$$

On the boundary $\partial\Omega$, homogeneous Dirichlet boundary conditions are prescribed. In our first numerical experiment we examine if the improved convergence properties of the modified scheme from [4] are preserved if $\boldsymbol{Q} = \boldsymbol{Q}(\boldsymbol{x}, c)$ depends on c. The advection-diffusion equation is solved for $D = 1$ on $\Omega = (0,1)^2$ with the water flux

$$\boldsymbol{Q}(x, y, c) = (-y, x)^{\top} \mathrm{e}^{-c}.$$

The source term F is choosen such that the analytical solution of the equation is given by $c(x, y) = x(1 - x)y(1 - y)$ which satisfies homogeneous Dirichlet boundary conditions. The coarsest mesh $\mathscr{T}_h = \{T_1, T_2\}$ with $T_1 = \mathrm{conv}\{(0,0), (1,0), (0,1)\}$ and $T_2 = \mathrm{conv}\{(1,0), (1,1), (0,1)\}$ is subject to a uniform refinement and the L^2 errors and experimental convergence rates $\alpha_k = \log(e_{k-1}/e_k)/\log(2)$ with e_k denoting the error on refinement level k are listed in Table 1. Here, by $\tilde{c}$ we denote the piecewise linear reconstruction of the concentrations in the local Crouzeix-Raviart space using the Lagrange multipliers; cf. [1]. For the reconstructed concentrations and the flux variable, second order convergence is clearly obtained in the numerical experiment. Obviously, the modified scheme works well also in this nonlinear case where $\boldsymbol{Q} = \boldsymbol{Q}(\boldsymbol{x}, c)$ depends on c. To obtain the second order flux approximation, it is crucial to discretize the nonlinearity in c using the Lagrange multipliers instead of the cellwise constant values c_K^n.

Table 2 Calculated errors and experimental convergence rates in the presence of reactions and nonlinear coefficient functions Θ and $\boldsymbol{Q}$ using the modified BDM_1 scheme

k	$\|c_1 - \tilde{c}_{1,h}\|$	α_k	$\|c_2 - \tilde{c}_{2,h}\|$	α_k	$\|\boldsymbol{q}_1 - \boldsymbol{q}_{1,h}\|$	α_k	$\|\boldsymbol{q}_2 - \boldsymbol{q}_{2,h}\|$	α_k
0	6.256e − 2		9.206e − 2		2.299e − 1		2.155e − 1	
1	1.643e − 2	1.93	2.317e − 2	1.99	6.565e − 2	1.81	6.098e − 2	1.82
2	4.160e − 3	1.98	5.800e − 3	2.00	1.733e − 2	1.92	1.603e − 2	1.93
3	1.040e − 3	2.00	1.453e − 3	2.00	4.460e − 3	1.96	4.121e − 3	1.96
4	2.568e − 4	2.02	3.659e − 4	1.99	1.136e − 3	1.97	1.055e − 3	1.97
5	6.158e − 5	2.06	9.411e − 5	1.96	2.866e − 4	1.99	2.718e − 4	1.96
6	1.666e − 5	1.89	2.632e − 5	1.84	7.619e − 5	1.91	7.319e − 5	1.89

3.2 Transport of Two Mixing and Reacting Species

In the following numerical example, the transport of two reacting and mixing substances in the domain $\Omega = (0, 2) \times (0, 3)$ is considered which is a modification of a test problem from [9]. The governing equations are

$$\partial_t(\Theta(c_1, c_2)c_i) - \nabla \cdot (\boldsymbol{D}_i \nabla c_i - \boldsymbol{Q}(c_1, c_2)c_i) = -\Theta(c_1, c_2)R_i(c_1, c_2) + F_i, \quad i = 1, 2,$$

where $R_i = \alpha_i\, c_1 c_2^2$ with $\alpha_1 = 1,\ \alpha_2 = 2$, $\boldsymbol{D}_1 = \boldsymbol{D}_2 = \boldsymbol{I}$. The saturation and the water flow $\boldsymbol{Q}$ are nonlinear functions of the concentrations, given by

$$\Theta(c_1, c_2) = e^{-\frac{1}{2}(c_1+c_2)}, \quad \boldsymbol{Q}(x, y, c_1, c_2) = (-y, x)^\top \mathrm{e}^{-(c_1+c_2)}.$$

The source terms F_i are choosen such that the analytical solutions for c_1 and c_2 are $c_1(x, y, t) = x(2.0 - x)y^3 \exp(-0.1t)/27$, $c_2(x, y, t) = n(x - 1.0)^2 y^2 \exp(-0.1t)/9$. The final simulation time is $T = 1$ and since only the spatial discretization error is of interest, the time step size $\tau^n = 0.001$ is chosen sufficiently small. The coarsest grid consisting of six triangles is refined uniformly and the computational results are summarized in Table 2. We clearly observe that in this numerical example a second-order accurate approximation of the flux variable is obtained using the method introduced above, whereas only suboptimal first order convergence was obtained for this problem with the classical scheme.

References

1. D. N. Arnold and F. Brezzi. Mixed and nonconforming finite element methods: implementation, postprocessing and error estimates. *RAIRO Modél. Math. Anal. Numér.*, 19(1):7–32, 1985.
2. M. Bause. Higher and lowest order mixed finite element approximation of subsurface flow problems with solutions of weak regularity. *Adv. Water Resour.*, 31:370–382, 2007.

3. F. Brezzi and M. Fortin. *Mixed and Hybrid Finite Element Methods*. Springer, New York, 1991.
4. F. Brunner, F.A. Radu, M. Bause, and P. Knabner. Optimal order convergence of a modified BDM_1 mixed finite element scheme for reactive transport in porous media. *Adv. Water Resour.*, 35:163–171, 2012.
5. Alan Demlow. Suboptimal and optimal convergence in mixed finite element methods. *SIAM J. Numer. Anal.*, 39:1938–1953, 2002.
6. R.H.W John, P. Porta, and Y. Vassilevski. Computational issues related to iterative coupling of subsurface and channel flows. *Calcolo*, 44:1–20, 2007.
7. P. Knabner, A. Prechtel, S. Bitterlich, R. Isa-Teran, and E. Schneid. Influence of surfactants on spreading of contaminants and soil remediation. In *Mathematics - Key Technology for the Future(Eds.) Jger W. and Krebs H.-J., Springer-Verlag, Berlin*, pages 152–161, 2003.
8. M. Ohlberger and C. Rohde. Adaptive finite volume approximations of weakly coupled convection dominated parabolic systems. *IMA J. Numer. Anal.*, 22:253–280, 2002.
9. F. A. Radu, M. Bause, A. Prechtel, and S. Attinger. A mixed hybrid finite element discretization scheme for reactive transport in porous media. In *Numerical Mathematics and Advanced Applications, Proceedings of ENUMATH 2007, the 7th European Conference on Numerical Mathematics and Advanced Applications*, pages 479–486, 2008.
10. F. A. Radu, N. Suciu, J. Hoffmann, A. Vogel, O. Kolditz, C.-H. Park, and S. Attinger. Accuracy of numerical simulations of contaminant transport in heterogeneous aquifers: A comparative study. *Adv. Water. Resour.*, 34:47–61, 2011.
11. C. E. Renshaw, G. D. Zynda, and J. C. Fountain. Permeability reductions induced by sorption of surfactant. *Water Resour. Res.*, 33:371–378, 1997.
12. B. Riviere and I. Yotov. Locally conservative coupling of stokes and darcy flows. *SIAM J. Numer. Anal.*, 42:1959–1977, 2005.
13. J. E. Smith and R. W. Gillham. Effects of solute concentration-dependent surface tension on unsaturated flow: Laboratory and column experiments. *Water Resour. Res.*, 35:973–982, 1999.
14. M. Vohralik. A posteriori error estimates for lowest-order mixed finite element discretizations of convection-diffusion-reaction equations. *SIAM J. Numer. Anal.*, 45:1570–1599, 2007.

A Robust Numerical Method for a Singularly Perturbed Parabolic Convection-Diffusion Problem with a Degenerating Convective Term and a Discontinuous Right-Hand Side

C. Clavero, J.L. Gracia, G.I. Shishkin, and L.P. Shishkina

Abstract In this paper we consider the efficient numerical approximation of a singularly perturbed parabolic convection-diffusion problem having a convective term which degenerates inside the domain, in the case that the right-hand side of the differential equation is discontinuous on the degeneration line. For small values of the diffusion parameter ε^2 ($\varepsilon \in (0,1]$), in general, the exact solution has an interior layer in a neighborhood of the degeneration line. We construct a classical finite difference scheme combining the implicit Euler method in time, defined on a uniform mesh, and the first order upwind scheme in space, defined on a piecewise-uniform grid condensing in a neighborhood of the interior layer. Then, the method is an ε-uniformly convergent scheme of first order in time and almost first order in space. We show the numerical results for a test problem, confirming in practice the theoretical results.

1 Introduction

On the set $\overline{G}$ with the boundary S, $\overline{G} = G \bigcup S$, $G = D \times (0,T]$, $D = (-d,d)$, we consider the initial-boundary value problem

$$L\,u(x,t) \equiv \left\{ \varepsilon^2 \frac{\partial^2}{\partial x^2} + x \frac{\partial}{\partial x} - \frac{\partial}{\partial t} - 1 \right\} u(x,t) = f(x,t), \quad (x,t) \in G \setminus S^{\pm}, \tag{1}$$

C. Clavero (✉) · J.L. Gracia
School of Engineering and Architecture, University of Zaragoza, Saragossa, Spain
e-mail: clavero@unizar.es; jlgracia@unizar.es

G.I. Shishkin · L.P. Shishkina
Institute of Mathematics and Mechanics, Ural Branch of Russian Academy of Sciences, Ekaterinburg, Russia
e-mail: shishkin@imm.uran.ru; lida@convex.ru

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_28, © Springer-Verlag Berlin Heidelberg 2013

where we assume that the function $f(x,t)$ is continuous on $\overline{G}^+$ and $\overline{G}^-$, where $\overline{G}^- = [-d,\,0]\times[0,\,T]$, $\overline{G}^+ = [0,\,d]\times[0,\,T]$, and it has a discontinuity of the first kind on the set $S^\pm = \{x=0\}\times(0,T]$, which is denoted by

$$[f(x^*,t)] \equiv f(x^*+0,t) - f(x^*-0,t) \neq 0, \quad (x^*,t)\in S^\pm. \tag{2}$$

On the set $S^\pm$ we impose the following continuity condition for the first-order derivative in x:

$$l^\pm u(x,t) \equiv \varepsilon\left[\frac{\partial}{\partial x}u(x+0,\,t) - \frac{\partial}{\partial x}u(x-0,\,t)\right] = 0, \quad (x,t)\in S^\pm. \tag{3}$$

On the boundary S, a Dirichlet boundary condition is prescribed

$$u(x,t) = \varphi(x,t), \quad (x,t)\in S. \tag{4}$$

The coefficient multiplying the first-order derivative in x in the differential equation (1) corresponds to the convective flux, which vanishes on $S^\pm$.

A function $u(x,t)$ is a solution of the initial-boundary value problem (1)–(4), which is continuous on $\overline{G}$ and it has continuous derivatives up to second order in x and continuous derivative in t, satisfying the differential equation (1) on $G\setminus S^\pm$. Continuity of the first-order derivative in x on the interface boundary $S^\pm$ in the conjunction condition (3), associates with the continuity of the diffusion flux across this boundary. Physical interpretations of the problem (1)–(4) are given in [1,11,14].

Characteristics of the reduced equation are tangent to the set $S^\pm$ on which the right-hand side has a discontinuity that causes appearance of an interior layer in the solution, when $\varepsilon\to 0$. However, characteristics of the reduced equation ($\varepsilon = 0$) enter into the domain $\overline{G}$, and they are not tangent to the lateral boundary; therefore, boundary layers in the problem (1)–(4) do not arise.

The analysis of ε-uniformly convergent numerical methods for singularly perturbed parabolic problems is well developed for the case when the problem data are sufficiently smooth and the convective term has constant sign (see, e.g., [4,7,9] and references therein). Nevertheless, there are few works in the case of discontinuous data and degenerating convective terms (see, e.g. [3,5,8,13]). Special schemes for problems with a convective term degenerating inside the domain and discontinuous data in equations were not considered earlier. Thus, our aim is to construct a ε-uniformly convergent finite difference scheme for the problem (1)–(4).

Below we denote by $S^L = S^l\bigcup S^r$ and $S_0 = S_0^+\bigcup S_0^-$ the lateral and lower parts of the boundary S, where $S_0 = [-d,\,d]\times\{t=0\}$, $S_0^+ = [0,\,d]\times\{t=0\}$, $S_0^- = [-d,\,0]\times\{t=0\}$, $S^l = \{x=-d\}\times(0,\,T]$, $S^r = \{x=d\}\times(0,\,T]$. Moreover, on the set $S^c = \overline{S}^L\bigcap S_0$ of corner points, and on the set $S^{\pm\,c} = \overline{S}^\pm\bigcap S_0$ of interior corner points, compatibility conditions are satisfied that guarantee the required smoothness of the solution in neighborhoods of these points.

2 Asymptotic Behavior of the Exact Solution

In this section we give the outlines of the analysis proving appropriate bounds of the solution of the continuous problem, which are used to prove the uniform convergence of the finite difference scheme defined in next section. A more extended version of these results, containing the theoretical and numerical analysis of the uniform convergence on uniform and special condensing meshes, is in preparation (see [2]). For the problem (1)–(4), the following comparison theorem holds (similar to one established in [10]).

Theorem 1. *Let the functions $u^1(x,t)$, $u^2(x,t)$ satisfy the relations*

$$\begin{aligned} L\,u^1(x,\,t) &\le L\,u^2(x,t), \quad (x,\,t) \in G \setminus S^{\pm},\\ l^{\pm}\,u^1(x,\,t) &\le l^{\pm}\,u^2(x,\,t), \quad (x,\,t) \in S^{\pm},\\ u^1(x,\,t) &\ge u^2(x,t), \quad (x,\,t) \in S. \end{aligned}$$

Then, $u^1(x,t) \ge u^2(x,t)$, $(x,\,t) \in \overline{G}$.

From Theorem 1, the bound $|u(x,\,t)| \le M$, $(x,\,t) \in \overline{G}$ follows. Classical theory (see [6]) permits to prove the following theorem, giving crude bounds for the derivatives.

Theorem 2. *We assume that the data of the initial-boundary value problem (1)–(4) satisfy $f \in C^{l_1-2,\,(l_1-2)/2}(\overline{G}^{+}) \cap C^{l_1-2,\,(l_1-2)/2}(\overline{G}^{-})$, $\varphi \in C^{l_1}(S_0^{+}) \cap C^{l_1}(S_0^{-}) \cap C^{l_1/2}(\overline{S}^{L})$ for $l_1 = l_0 + \alpha$, $\alpha > 0$, with $l_0 > 0$ is even. Then, the solution of (1)–(4) satisfies*

$$\left| \frac{\partial^{k+k_0}}{\partial x^k\,\partial t^{k_0}}\,u(x,t) \right| \le M\varepsilon^{-2k}, \quad (x,t) \in \overline{G}^{+} \cup \overline{G}^{-}, \quad k + 2k_0 \le l_0.$$

To improve these crude estimates, we decompose the solution as $u(x,\,t) = U^{+}(x,\,t) + V^{+}(x,\,t)$, $(x,\,t) \in \overline{G}^{+}$, $u(x,\,t) = U^{-}(x,\,t) + V^{-}(x,\,t)$, $(x,\,t) \in \overline{G}^{-}$, where $U^{+}(x,\,t)$, $U^{-}(x,\,t)$ and $V^{+}(x,\,t)$, $V^{-}(x,\,t)$ are the regular and the singular components of the solution, respectively. The function $U^{+}(x,t)$, $(x,\,t) \in \overline{G}^{+}$ is the restriction to $\overline{G}^{+}$ of the function $U^{+e}(x,\,t)$, $(x,\,t) \in \overline{G}^{+e}$, where $U^{+e}(x,t)$, $(x,\,t) \in \overline{G}^{e}$ is the solution of the initial-boundary value problem

$$L^e U^{+e}(x,t) = f^{+e}(x,t),\ (x,t) \in G^{+e}, \quad U^{+e}(x,t) = \varphi^{+e}(x,t),\ (x,t) \in S^{+e},$$

where we take $G^{+e} = G$, and $f^{+e}(x,t)$ and $\varphi^{+e}(x,t)$ are smooth extensions of $f(x,t)$ and $\varphi(x,t)$, respectively.

Then, for the function $U^+(x,t)$ on $\overline{G}^+$, it is possible to prove the estimates

$$\left|\frac{\partial^{k+k_0}}{\partial x^k \partial t^{k_0}} U^+(x,t)\right| \leq M[1+\varepsilon^{2(2-k-k_0)}], \quad (x,t)\in \overline{G}^+, \quad k+2k_0 \leq l_0-2.$$

On the other hand, the function $V^+(x,t)$, $(x,t)\in \overline{G}^+$ is the solution of the problem

$$\begin{aligned} L\, V^+(x,t) &= 0, && (x,t)\in G^+,\\ V^+(x,t) &= \varphi_{V^+}(x,t), && (x,t)\in S^{\pm},\\ V^+(x,t) &= 0, && (x,t)\in S_0^+ \cup S^r, \end{aligned}$$

where $\varphi_{V^+}(x,t) = \varphi_u(x,t) - U^+(x,t)$, $(x,t)\in S^{\pm}$. Then, a detailed analysis (see [2]) based on a standard comparison theorem with the majorant function

$$W(x,t) = M\left[1+\varepsilon^{2(2-k_0)}\right]\left\{1-\Phi\left(m_0^{1/2}\frac{x}{\varepsilon}\right)\right\}, \quad (x,t)\in \overline{G}^+,$$

where $\Phi(z) = (2/\sqrt{\pi})\int_0^z e^{-\zeta^2} d\,\zeta$ is the error integral (see, e.g., [14], Chap. 3), and m_0 is an arbitrary constant satisfying the inequality $m_0 \leq 1/2$, gives the estimate

$$\left|\frac{\partial^{k+k_0}}{\partial x^k \partial t^{k_0}} V^+(x,t)\right| \leq M\left[\varepsilon^{-k}+\varepsilon^{2(2-k_0)-k}\right]\exp^{-m_1\frac{x^2}{\varepsilon^2}}, \ (x,t)\in \overline{G}^+, \ k+2k_0 \leq l_0-2,$$

with $m_1 < 1/2$. In a similar way, for $U^-(x,t)$ and $V^-(x,t)$ on the set $\overline{G^-}$ it holds

$$\begin{aligned} &\left|\frac{\partial^{k+k_0}}{\partial x^k \partial t^{k_0}} U^-(x,t)\right| \leq M\left[1+\varepsilon^{2(2-k-k_0)}\right], \ (x,t)\in \overline{G^-}, \ k+2k_0 \leq l_0-2\\ &\left|\frac{\partial^{k+k_0}}{\partial x^k \partial t^{k_0}} V^-(x,t)\right| \leq M\left[\varepsilon^{-k}+\varepsilon^{2(2-k_0)-k}\right]\exp^{-m_1\frac{x^2}{\varepsilon^2}}, \ (x,t)\in \overline{G^-}, \ k+2k_0 \leq l_0-2. \end{aligned}$$

From this estimates, the following theorem follows.

Theorem 3. *Let for the data of the initial-boundary value problem (1)–(4) the conditions* $f \in C^{l_1}(\overline{G}^+) \cap C^{l_1}(\overline{G^-})$, $\varphi \in C^{l_1}(S_0^+) \cap C^{l_1}(S_0^-) \cap C^{l_1}(\overline{S}^L)$, *where* $l_1 = l_0+\alpha$ *and* $l_0 = 6$. *Then, the following estimates are valid:*

$$\left|\begin{array}{ll}
\dfrac{\partial^{k+k_0}}{\partial x^k\,\partial t^{k_0}}\,U^+(x,t)\right| \le M\left[1+\varepsilon^{2(2-k-k_0)}\right], & (x,t)\in\overline{G}^+;\\
\left|\dfrac{\partial^{k+k_0}}{\partial x^k\,\partial t^{k_0}}\,V^+(x,t)\right| \le M\left[\varepsilon^{-k}+\varepsilon^{2(2-k_0)-k}\right]\exp^{-m_1\frac{x^2}{\varepsilon^2}}, & (x,t)\in\overline{G}^+;\\
\left|\dfrac{\partial^{k+k_0}}{\partial x^k\,\partial t^{k_0}}\,U^-(x,t)\right| \le M\left[1+\varepsilon^{2(2-k-k_0)}\right], & (x,t)\in\overline{G}^-;\\
\left|\dfrac{\partial^{k+k_0}}{\partial x^k\,\partial t^{k_0}}\,V^-(x,t)\right| \le M\left[\varepsilon^{-k}+\varepsilon^{2(2-k_0)-k}\right]\exp^{-m_1\frac{x^2}{\varepsilon^2}}, & (x,t)\in\overline{G}^-,
\end{array}\right. \tag{5}$$

for $k+2k_0\le 4$ *and* $m_1<1/2$.

Moreover, for the components $V^+(x,\,t)$ and $V^-(x,\,t)$ we have also the estimates

$$\left|V^+(x,\,t)\right|\le M\,\exp(-m\varepsilon^{-1}x),\ (x,\,t)\in\overline{G}^+,\ \left|V^-(x,\,t)\right|\le M\,\exp(m\varepsilon^{-1}x),\ (x,\,t)\in\overline{G}^-, \tag{6}$$

where m is an arbitrary constant satisfying the condition $m\le 1$.

3 A Finite Difference Scheme on a Special Mesh

From estimates in Theorem 3 we know that an interior layer appears in the exact solution. Then, we introduce a grid condensed in a neighborhood of the interior layer (see, e.g., [11, 12] and the bibliography therein), $\overline{G}_h^s\equiv\overline{\omega}^s\times\overline{\omega}_0^u$, where $\overline{\omega}_0^u$ is a uniform mesh in time and $\overline{\omega}^s$ is a piecewise-uniform mesh constructed in the following way. We divide the interval $[-d,\,d]$ into three parts $[-d,\,-\sigma]$, $[-\sigma,\,\sigma]$ and $[\sigma,\,d]$, where the transition parameter σ is given by

$$\sigma=\sigma(\varepsilon,\,N)=\min\left[4^{-1}\,d,\,m_2^{-1}\,\varepsilon\,\ln N\right], \tag{7}$$

with m_2 an arbitrary number from the interval $(0,m)$ and m is given in (6).Outside the set $S_h^{\pm}=S^{\pm}\cap G_h^s$, we approximate the problem (1)–(4) by the finite difference scheme

$$\Lambda\,z(x,t)\equiv\left\{\varepsilon^2\,\delta_{\widehat{xx}}+x\,\delta_x^*-\delta_{\bar{t}}-1\right\}z(x,t)=f(x,t),\quad (x,t)\in G_h^s\setminus S^{\pm}, \tag{8}$$

where

$$\delta_x^*\,z(x,t)=\begin{cases}\delta_x\,z(x,t), & x>0,\\ \delta_{\bar{x}}\,z(x,t), & x<0,\end{cases}$$

is the monotone approximation of the derivative $\frac{\partial}{\partial x} u(x,t)$ in the differential equation, $\delta_{\widehat{\overline{x}x}} z(x,t)$ is the central second-order difference derivative on a nonuniform grid, $\delta_x z(x,t)$ and $\delta_{\overline{x}} z(x,t)$ are the forward and backward, respectively, finite differences for the space variable, and $\delta_{\overline{t}} z(x,t)$ is the backward finite difference for the time variable. Also we impose the continuity of the discrete flux in $S_h^{\pm}$, i.e.,

$$\Lambda^{\pm} z(x,t) \equiv \varepsilon \left[\delta_x z(x,t) - \delta_{\overline{x}} z(x,t)\right] = 0, \quad (x,\, t) \in S_h^{\pm}, \tag{9}$$

and on the boundary we have the condition $z(x,t) = \varphi(x,t)$, $(x,\, t) \in S_h = S \cap G_h^s$.

Using a standard technique (see [13]), it is possible to deduce appropriate bounds for the truncation error associated to the finite difference scheme (8) and (9) on the grid $\overline{G}_h^s$, and therefore, using the discrete minimum principle, the estimate

$$|u(x,t) - z(x,t)| \leq M \left\{ N^{-1} \min \left[\varepsilon^{-1}, \ln N\right] + N_0^{-1} \right\}, \quad (x,t) \in \overline{G}_h^s, \tag{10}$$

follows. Also we can prove the ε-uniform estimate

$$|u(x,t) - z(x,t)| \leq M \left[N^{-1} \ln N + N_0^{-1}\right], \quad (x,t) \in \overline{G}_h^s. \tag{11}$$

showing that the numerical scheme is ε-uniformly convergent with the first order in time and almost first order in space.

4 Numerical Results

In this section we show the numerical results obtained for a test problem of type (1)–(4), when the right-hand side is given by

$$f(x,t) = \begin{cases} -t^2(\cos(\pi x) + e^x), & \text{if } x > 0, \\ 0, & \text{if } x < 0. \end{cases} \tag{12}$$

Plot of the discrete solution of problem (1) under the condition (12) is given in Fig. 1 on the piecewise uniform mesh (7) with $m_2 = 1/2$; from it we see that the scheme on this mesh resolves the interior layer.

To estimate the numerical errors on both the uniform and the piecewise-uniform mesh, we use a variant of the double mesh principle, and then the approximated error $D_{i,n}^{\varepsilon,N,N_0}$ is defined by

$$D_{i,n}^{\varepsilon,N,N_0} = \left| U_{i,n}^{\varepsilon,N,N_0} - U_{i,n}^{\varepsilon,2N,2N_0} \right|,$$

computed on the nodes $(x^i, t_n) \in \overline{G}_h^s$, where $U_{i,n}^{\varepsilon,N,N_0}$ is the numerical solution obtained by using the constant time step $\tau = 1/N_0$ and $(N+1)$ points in the

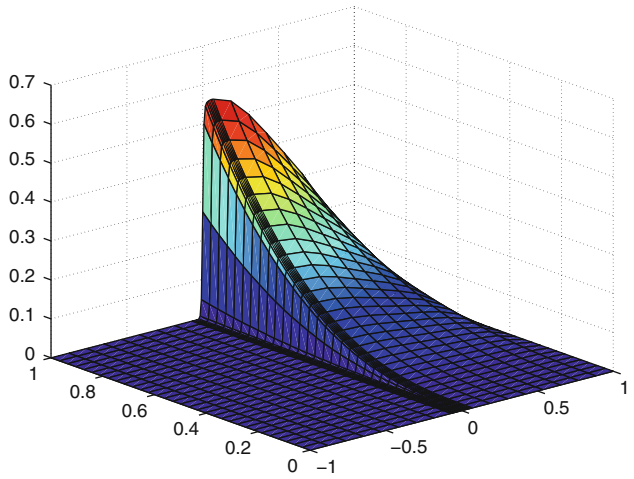


Fig. 1 Numerical solution of problem (1), (12) for $\varepsilon = 10^{-2}$ with $N = N_0 = 32$ on the mesh (7)

Table 1 Maximum errors and orders of convergence on a uniform mesh

ε	$N = 32$	$N = 64$	$N = 128$	$N = 256$	$N = 512$	$N = 1,024$	$N = 2,048$
2^{-1}	$0.155E-2$	$0.843E-3$	$0.439E-3$	$0.224E-3$	$0.113E-3$	$0.567E-4$	$0.284E-4$
	0.880	0.942	0.972	0.986	0.993	0.997	
2^{-2}	$0.371E-2$	$0.195E-2$	$0.100E-2$	$0.510E-3$	$0.257E-3$	$0.129E-3$	$0.645E-4$
	0.923	0.959	0.979	0.990	0.995	0.997	
2^{-3}	$0.663E-2$	$0.365E-2$	$0.192E-2$	$0.987E-3$	$0.500E-3$	$0.251E-3$	$0.126E-3$
	0.859	0.925	0.964	0.982	0.991	0.995	
2^{-4}	$0.103E-1$	$0.576E-2$	$0.256E-2$	$0.121E-2$	$0.593E-3$	$0.298E-3$	$0.149E-3$
	0.841	1.170	1.078	1.033	0.993	0.995	
2^{-5}	$0.802E-2$	$0.965E-2$	$0.509E-2$	$0.222E-2$	$0.106E-2$	$0.517E-3$	$0.256E-3$
	−0.268	0.923	1.198	1.072	1.032	1.012	
2^{-6}	$0.809E-2$	$0.728E-2$	$0.922E-2$	$0.469E-2$	$0.203E-2$	$0.981E-3$	$0.480E-3$
	0.152	−0.340	0.974	1.209	1.048	1.032	
2^{-7}	$0.811E-2$	$0.456E-2$	$0.708E-2$	$0.898E-2$	$0.448E-2$	$0.195E-2$	$0.942E-3$
	0.830	−0.634	−0.342	1.004	1.197	1.051	
2^{-8}	$0.812E-2$	$0.457E-2$	$0.265E-2$	$0.697E-2$	$0.885E-2$	$0.436E-2$	$0.192E-2$
	0.830	0.784	−1.395	−0.344	1.020	1.186	
2^{-9}	$0.812E-2$	$0.457E-2$	$0.240E-2$	$0.262E-2$	$0.692E-2$	$0.878E-2$	$0.431E-2$
	0.830	0.926	−0.122	−1.403	−0.345	1.028	
2^{-10}	$0.812E-2$	$0.457E-2$	$0.240E-2$	$0.123E-2$	$0.260E-2$	$0.689E-2$	$0.875E-2$
	0.830	0.926	0.962	−1.074	−1.408	−0.345	
...	...	...	...	...	...	...	...
2^{-15}	$0.812E-2$	$0.457E-2$	$0.240E-2$	$0.123E-2$	$0.625E-3$	$0.315E-3$	$0.158E-3$
	0.830	0.926	0.962	0.981	0.991	0.995	
D^{N,N_0}	$0.103E-1$	$0.965E-2$	$0.922E-2$	$0.898E-2$	$0.885E-2$	$0.878E-2$	$0.875E-2$
q_{uni}	0.096	0.067	0.038	0.021	0.011	0.005	

Table 2 Maximum errors and uniform orders of convergence on the piecewise–uniform mesh

ε	$N = 32$	$N = 64$	$N = 128$	$N = 256$	$N = 512$	$N = 1,024$	$N = 2,048$
2^{-1}	$0.155E-2$	$0.843E-3$	$0.439E-3$	$0.224E-3$	$0.113E-3$	$0.567E-4$	$0.284E-4$
	0.880	0.942	0.972	0.986	0.993	0.997	
2^{-2}	$0.371E-02$	$0.195E-02$	$0.100E-02$	$0.510E-03$	$0.257E-03$	$0.129E-03$	$0.645E-04$
	0.923	0.959	0.979	0.990	0.995	0.997	
2^{-3}	$0.663E-02$	$0.365E-02$	$0.192E-02$	$0.987E-03$	$0.500E-03$	$0.251E-03$	$0.126E-03$
	0.859	0.925	0.964	0.982	0.991	0.995	
2^{-4}	$0.990E-02$	$0.576E-02$	$0.256E-02$	$0.121E-02$	$0.593E-03$	$0.298E-03$	$0.149E-03$
	0.781	1.170	1.078	1.033	0.993	0.995	
2^{-5}	$0.118E-01$	$0.653E-02$	$0.331E-02$	$0.160E-02$	$0.844E-03$	$0.455E-03$	$0.246E-03$
	0.847	0.982	1.043	0.927	0.891	0.890	
2^{-6}	$0.132E-01$	$0.763E-02$	$0.405E-02$	$0.206E-02$	$0.102E-02$	$0.502E-03$	$0.245E-03$
	0.793	0.915	0.976	1.010	1.026	1.035	
2^{-7}	$0.138E-01$	$0.816E-02$	$0.441E-02$	$0.228E-02$	$0.115E-02$	$0.575E-03$	$0.285E-03$
	0.758	0.887	0.953	0.985	1.002	1.010	
2^{-8}	$0.140E-01$	$0.839E-02$	$0.458E-02$	$0.239E-02$	$0.122E-02$	$0.611E-03$	$0.305E-03$
	0.743	0.873	0.941	0.974	0.992	1.000	
2^{-9}	$0.141E-01$	$0.850E-02$	$0.466E-02$	$0.244E-02$	$0.125E-02$	$0.629E-03$	$0.316E-03$
	0.735	0.866	0.932	0.970	0.987	0.996	
2^{-10}	$0.142E-01$	$0.855E-02$	$0.470E-02$	$0.247E-02$	$0.126E-02$	$0.638E-03$	$0.321E-03$
	0.732	0.862	0.929	0.968	0.985	0.993	
...	...	...	...	...	...	...	...
2^{-15}	$0.142E-01$	$0.859E-02$	$0.475E-02$	$0.250E-02$	$0.128E-02$	$0.647E-03$	$0.325E-03$
	0.729	0.856	0.928	0.965	0.983	0.991	
D^{N,N_0}	$0.142E-01$	$0.859E-02$	$0.475E-02$	$0.250E-02$	$0.128E-02$	$0.647E-03$	$0.325E-03$
q_{uni}	0.729	0.856	0.928	0.965	0.983	0.991	

spatial mesh, and $U_{i,n}^{\varepsilon,2N,2N_0}$ is computed using $\tau/2$ as a time step and $(2N+1)$ points in the spatial mesh, but with the same transition parameter as in the original mesh. For each fixed value of ε, the maximum global errors D^{ε,N,N_0} are estimated by $D^{\varepsilon,N,N_0} = \max_{i,n} D_{i,n}^{\varepsilon,N,N_0}$, and the numerical orders of convergence q are given by

$$q = q(\varepsilon, N, N_0) = \log\left(D^{\varepsilon,N,N_0}/D^{\varepsilon,2N,2N_0}\right)/\log 2.$$

From these values, the computed errors D^{N,N_0} and the computed orders of uniform convergence q_{uni} are given by

$$D^{N,N_0} = \max_{\varepsilon} D^{\varepsilon,N,N_0}, \quad q_{uni} = q_{uni}(N, N_0) = \log\left(D^{N,N_0}/D^{2N,2N_0}\right)/\log 2.$$

Table 1 displays the errors and the orders of convergence for the test problem on a uniform mesh. From it we can observe that the method is not ε-uniform convergent.

Table 2 displays the errors and the orders of convergence on the piecewise uniform mesh taking $m_2 = 1/2$, for the same values of ε as before. From it we clearly observe the ε-uniform convergence of the scheme and the first convergence order of the method.

Acknowledgements This research was partially supported by the project MEC/FEDER MTM 2010-16917 and the Diputación General de Aragón and also by the Russian Foundation for Basic Research under grant 10-01-00726.

References

1. Budak, B.M., Samarskii, A.A., Tikhonov, A.N.: Problems of Mathematical Physics. Nauka, Moscow (1980) (in Russian).
2. Clavero, C., Gracia, J.L., Shishkin, G.I, Shishkina, L.P.: Grid approximation of a singularly perturbed parabolic equation with degenerating convective term and discontinuous right-hand side, in preparation.
3. Dunne, R.K., O'Riordan, E., Shishkin, G.I.: A fitted mesh method for a class of singularly perturbed parabolic problems with a boundary turning point. Comput. Meth. Appl. Math., **3**, 361–372 (2003).
4. Farrell, P.A., Hegarty, A.F., Miller, J.J.H., O'Riordan, E., Shishkin, G.I.: Robust Computational Techniques for Boundary Layers. Applied Mathematics, 16. Chapman and Hall/CRC (2000).
5. Hemker, P.W., Shishkin, G.I.: Discrete approximation of singularly perturbed parabolic PDEs with a discontinuous initial condition, Comput. Fluid Dynamics J., **2**, 375–392 (1994).
6. Ladyzhenskaya, O.A., Solonnikov, V.A., Uraltseva, N.N.: Linear and Quasilinear Equations of Parabolic Type. Nauka, Moscow (1967) (in Russian). Translations of Mathematical Monographs, **23**, American Mathematical Society, Providence, R.I. (1967).
7. Miller, J.J.H., O'Riordan, E., Shishkin, G.I.: Fitted Numerical Methods for Singular Perturbation Problems. World Scientific, Singapore (1996).
8. O'Riordan, E., Shishkin, G.I.: Singularly perturbed parabolic problems with non-smooth data, J. Comp. Appl. Math., **166**, 233–245 (2004).
9. Roos, H.-G., Stynes, M., Tobiska, L: Numerical Methods for Singularly Perturbed Differential Equations. Convection-Diffusion-Reaction and Flow Problems, second edition, Springer Series in Computational Mathematics, **24**, Springer-Verlag, Berlin (2008).
10. Shishkin, G.I.: Grid Approximations of Singularly Perturbed Elliptic and Parabolic Equations. Ural Branch of RAS, Ekaterinburg, 1992 (in Russian).
11. Shishkin, G.I.: On numerical methods on adaptive meshes for a singularly perturbed reaction-diffusion equation with a moving concentrated source. Finite Difference Schemes, Lith. Acad. Sci., Inst. Math. Inform., Vilnius, 205–214 (2000).
12. Shishkin, G.I.: Grid approximation of a singularly perturbed parabolic equation on a composed domain with a moving interface containing a concentrated source. Comput. Maths. Math. Phys., **43**, 1738–1755 (2003).
13. Shishkin, G.I., Shishkina, L.P.: Difference Methods for Singular Perturbation Problems. Series: Monographs & Surveys in Pure & Applied Math. Chapman and Hall/CRC (2009).
14. Tikhonov, A.N., Samarskii, A.A.: Equations of Mathematical Physics. Dover Publications, Inc., New York (1990).

Finite Element Methods with Artificial Diffusion for Hamilton-Jacobi-Bellman Equations

M. Jensen and I. Smears

Abstract In this short note we investigate the numerical performance of the method of artificial diffusion for second-order fully nonlinear Hamilton-Jacobi-Bellman equations. The method was proposed in (Jensen and Smears, On the convergence of finite element methods for Hamilton-Jacobi-Bellman equations, arxiv:1111.5423, 2011); where a framework of finite element methods for Hamilton-Jacobi-Bellman equations was studied theoretically. The numerical examples in this note study how the artificial diffusion is activated in regions of degeneracy, the effect of a locally selected diffusion parameter on the observed numerical dissipation and the solution of second-order fully nonlinear equations on irregular geometries.

1 Introduction

Hamilton-Jacobi-Bellman (HJB) equations, which describe the value function in the theory of optimal control, are fully nonlinear partial differential equations, which are of second-order if the underlying control problem is stochastic. One challenge arising in the numerical solution of these equations is the presence of spurious generalised solutions of the PDE which do not coincide with the value function. While these spurious solutions often possess the same regularity as the value function, they violate monotonicity properties exhibited by the value function. These properties lead to the concept of viscosity solutions.

M. Jensen (✉)
Durham University, Durham, UK
e-mail: m.p.j.jensen@durham.ac.uk

I. Smears
Oxford University, Oxford, UK
e-mail: iain.smears@maths.ox.ac.uk

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_29,

Regarding the numerical solution of HJB equations, we would like to highlight three approaches within the finite element methodology, which have been employed to ensure convergence to the value function. For a review of discrete Markov chain approximations before application of the dynamic programming principle we refer to [5]. For the method of vanishing moments we point to [3]. For the approach by Barles and Souganidis we cite the original source [1] and the recent adaption [4] by the authors to a finite element framework. A more comprehensive outline of the literature is in [2].

2 Numerical Method

Let Ω be a bounded Lipschitz domain in $\mathbb{R}^d$, $d \geq 2$. Let A be a compact metric space and

$$A \to C(\overline{\Omega}) \times C(\overline{\Omega}, \mathbb{R}^d) \times C(\overline{\Omega}) \times C(\overline{\Omega}),\ \alpha \mapsto (a^\alpha, b^\alpha, c^\alpha, d^\alpha)$$

be continuous. Consider the bounded linear operators of non-negative characteristic form

$$L^\alpha :\ H^2(\Omega) \cap H_0^1(\Omega) \to L^2(\Omega),\ w \mapsto -a^\alpha\, \Delta w + b^\alpha \cdot \nabla w + c^\alpha\, w$$

where α belongs to A. Then

$$\sup_{\alpha \in A} \| (a^\alpha, b^\alpha, c^\alpha, d^\alpha) \|_{C(\overline{\Omega}) \times C(\overline{\Omega}, \mathbb{R}^d) \times C(\overline{\Omega}) \times C(\overline{\Omega})} + \sup_{\alpha \in A} \| L^\alpha \|_{C^2(\overline{\Omega}) \to C(\overline{\Omega})} < \infty. \quad (1)$$

We assume that the final-time boundary data $v_T \in C(\overline{\Omega})$ is non-negative: $v_T \geq 0$ on $\overline{\Omega}$. For smooth w let $Hw := \sup_\alpha (L^\alpha w - d^\alpha)$, where the supremum is applied pointwise. The HJB equation considered is

$$-v_t + Hv = 0 \qquad \text{in } (0,T) \times \Omega, \quad (2a)$$

$$v = g \qquad \text{on } (0,T) \times \partial\Omega, \quad (2b)$$

$$v = v_T \qquad \text{on } \{T\} \times \overline{\Omega}. \quad (2c)$$

Let V_i, $i = 1, 2, \ldots$, be a sequence of piecewise linear shape-regular finite element spaces, whose underlying meshes are strictly acute. Let y_i^ℓ, $\ell = 1, \ldots, \dim V_i$, denote the nodes of the mesh with associated hat functions ϕ_i^ℓ. Set $\hat{\phi}_i^\ell := \phi_i^\ell / \| \phi_i^\ell \|_{L^1(\Omega)}$. The mesh size is denoted $(\Delta x)_i$. The set of time steps is $S_i := \{ s_i^k : k = 0, \ldots, T/h_i \}$. Let the ℓth entry of $d_i w(s_i^k, \cdot)$ be $(d_i w(s_i^k, \cdot))_\ell = (w(s_i^{k+1}, y_i^\ell) - w(s_i^k, y_i^\ell))/h_i$.

For each α and i find an approximate splitting $L^\alpha \approx E_i^\alpha + I_i^\alpha$ into linear operators

$$E_i^\alpha : \quad w \mapsto -\bar{a}_i^\alpha \, \Delta w + \bar{b}_i^\alpha \cdot \nabla w + \bar{c}_i^\alpha \, w,$$
$$I_i^\alpha : \quad w \mapsto -\bar{\bar{a}}_i^\alpha \, \Delta w + \bar{\bar{b}}_i^\alpha \cdot \nabla w + \bar{\bar{c}}_i^\alpha \, w,$$

of the form $a^\alpha = \tilde{a}_i^\alpha + \tilde{\tilde{a}}_i^\alpha$, $b^\alpha = \bar{b}_i^\alpha + \bar{\bar{b}}_i^\alpha$, $c^\alpha = \bar{c}_i^\alpha + \bar{\bar{c}}_i^\alpha$ and $d^\alpha = d_i^\alpha$. To impose monotonicity, select the artificial diffusion parameters $\bar{\nu}_i^{\alpha,\ell}$ and $\bar{\bar{\nu}}_i^{\alpha,\ell}$ as in [4] such that $\bar{a}_i^\alpha(y_i^\ell) \geq \max\{\tilde{a}_i^\alpha(y_i^\ell), \bar{\nu}_i^{\alpha,\ell}\}$ and $\bar{\bar{a}}_i^\alpha(y_i^\ell) \geq \max\{\tilde{\tilde{a}}_i^\alpha(y_i^\ell), \bar{\bar{\nu}}_i^{\alpha,\ell}\}$.

Define, for $w \in H^1(\Omega)$, $\ell \in \{1, \dots, N = \dim V_i^0\}$,

$$(\mathsf{E}_i^\alpha w)_\ell := \bar{a}_i^\alpha(y_i^\ell)\langle \nabla w, \nabla \hat{\phi}_i^\ell \rangle + \langle \bar{b}_i^\alpha \cdot \nabla w + \bar{c}_i^\alpha \, w, \hat{\phi}_i^\ell \rangle, \tag{3a}$$
$$(\mathsf{I}_i^\alpha w)_\ell := \bar{\bar{a}}_i^\alpha(y_i^\ell)\langle \nabla w, \nabla \hat{\phi}_i^\ell \rangle + \langle \bar{\bar{b}}_i^\alpha \cdot \nabla w + \bar{\bar{c}}_i^\alpha w, \hat{\phi}_i^\ell \rangle, \tag{3b}$$
$$(\mathsf{C}_i^\alpha)_\ell := \langle d_i^\alpha, \hat{\phi}_i^\ell \rangle. \tag{3c}$$

Obtain the numerical solution $v_i(T, \cdot) \in V_i$ by interpolation of v_T. Then $v_i(s_i^k, \cdot) \in V_i$ at time s_i^k is defined, inductively, by interpolating the boundary data and by

$$-d_i \, v_i(s_i^k, \cdot) + \sup_\alpha \big(\mathsf{E}_i^\alpha v_i(s_i^{k+1}, \cdot) + \mathsf{I}_i^\alpha v_i(s_i^k, \cdot) - \mathsf{C}_i^\alpha \big) = 0, \tag{4}$$

where the supremum is understood to be applied component-wise to the collection of vectors $\big\{ \mathsf{E}_i^\alpha v_i(s_i^{k+1}, \cdot) + \mathsf{I}_i^\alpha v_i(s_i^k, \cdot) - \mathsf{C}_i^\alpha \, : \, \alpha \in A \big\}$.

3 Selection of the Artificial Diffusion Parameter

That the diffusion coefficients $\bar{a}_i^\alpha$ and $\bar{\bar{a}}_i^\alpha$ in (3) are placed outside of the scalar product originates from the non-divergence form of the linear operators L^α of HJB operators [4]. This structure makes it straightforward to implement *local* artificial diffusion parameters $\bar{\nu}_i^{\alpha,\ell}$ and $\bar{\bar{\nu}}_i^{\alpha,\ell}$, that is to implement a dependence on the nodal position y_i^ℓ. Indeed a change of $\bar{\nu}_i^{\alpha,\ell}$ corresponds in the assembly of E_i^α to a scalar multiplication of the ℓth row of the stiffness matrix, see (3).

Changes in the optimal artificial diffusion coefficients arise from variations of the local mesh quality and the local mesh Péclet number. The parameters $\bar{\nu}_i^{\alpha,\ell}$ and $\bar{\bar{\nu}}_i^{\alpha,\ell}$ may be chosen by studying (35) in [4] or by examining the assembled matrices of the unstabilised operators. This is in particular simple for the E_i^α, since only the signs of off-diagonal entries need be corrected to impose a local monotonicity property, thus leading to an algorithm which can be executed row-by-row.

Notice that in contrast, for problems in divergence form, the stabilised diffusion coefficients $\tilde{a}_i^\alpha$ and $\tilde{\tilde{a}}_i^\alpha$ need to be determined in the whole domain and not just at nodal positions.

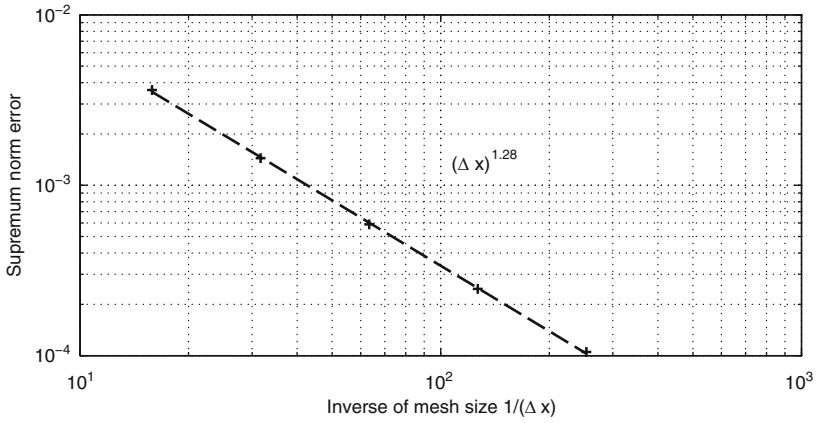


Fig. 1 Uniform error at the initial time $t = 0$ with constant ratio $h_i/(\Delta x)_i$

4 An Exact Solution and Convergence Rates

We consider a triangular spatial domain Ω with the vertices $(0, -1)$, $(\sqrt{3}/2, 1/2)$ and $(-\sqrt{3}/2, 1/2)$ and the HJB equation

$$-v_t - \frac{1}{2}\sqrt{\frac{x^2+y^2}{T-t+1}}\Delta v + \frac{1}{2}\frac{1}{\sqrt{T-t+1}}|\nabla v| = -\frac{1}{2}\frac{\sqrt{x^2+y^2}}{(T-t+1)^{3/2}}. \tag{5}$$

To see that Eq. (5) is a HJB equation, note that the Euclidean norm of the gradient satisfies

$$|\nabla v| = \sup\{\beta \cdot \nabla v \,:\, \beta \in \mathbb{R}^2 \text{ with } |\beta| = 1\}. \tag{6}$$

A calculation verifies that the function

$$v(x, y, t) = \exp\left(-\sqrt{\frac{x^2+y^2}{T-t+1}}\right) + \sqrt{\frac{x^2+y^2}{T-t+1}}$$

is an exact solution of (5), where boundary and final-time conditions are determined by interpolation of the exact solution v. The equation is to be solved on the time interval $(0, 1)$.

We consider a splitting which discretises the advection term explicitly with the minimum amount of diffusion needed for monotonicity. The remaining (linear) diffusion is incorporated in the implicit term, observing that this leads to a time-step

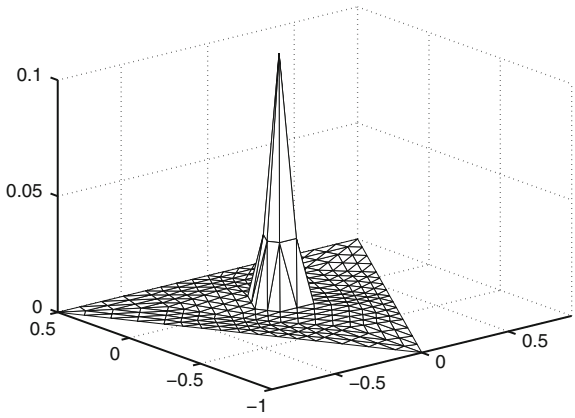


Fig. 2 The plot shows a locally-adapted choice of artificial diffusion over the triangular domain. The peak in artificial diffusion at the centre of the domain corresponds to the degeneracy of the differential operator at the origin

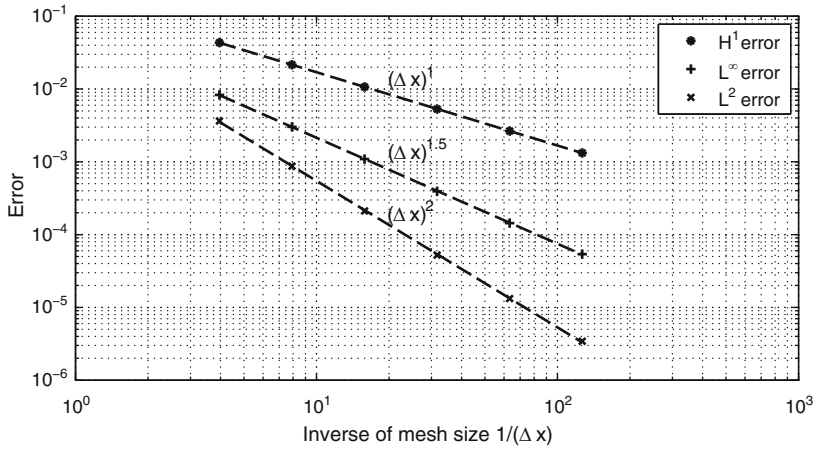


Fig. 3 Convergence rates of the numerical solution at the initial time $t = 0$ in the L^2-, L^∞- and H^1-norms with constant ratio $h_i/(\Delta x)_i^2$

restriction $h_i \lesssim (\Delta x)_i$. Figure 1 shows the supremum norm error with a constant ratio $h_i/(\Delta x)_i$, showing the uniform stability of the method in this setting.

Figure 2 illustrates a choice of artificial diffusion that is locally adapted to the Péclet number of a coarse mesh. It is seen that diffusion is only artificially introduced in a neighbourhood of the origin, which is where the operator becomes degenerate. Figure 3 illustrates the rates of convergence of the numerical solution $v_i(0,\cdot)$ in the L^2-, L^∞- and H^1-norms to the exact solution at the initial time, now using the constant ratio $h_i/(\Delta x)_i^2$.

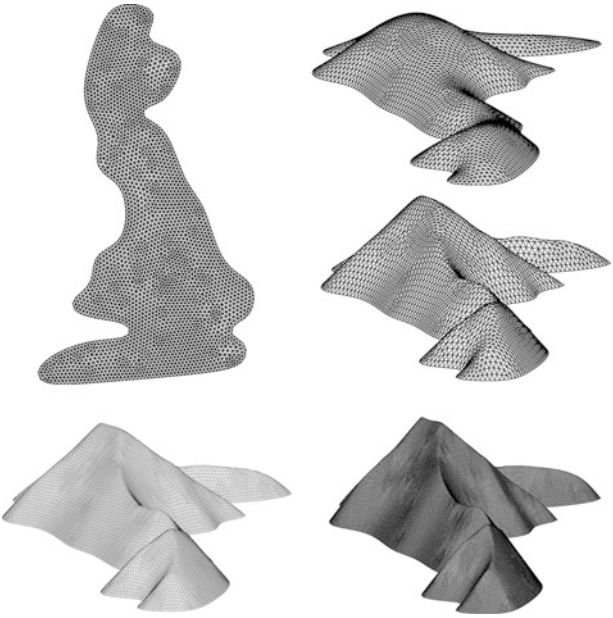


Fig. 4 The *top left plot* shows the original mesh for the domain. The *top right plot* shows the numerical solution with globally-chosen diffusion and below is a plot with locally-chosen artificial diffusion. The *bottom left plot* is the numerical solution from one uniform refinement of the mesh and the *bottom right plot* shows the solutions after two refinements

5 Eikonal Equation

The steady-state limit of the time-dependent eikonal equation,

$$-v_t + |\nabla v| = 1,$$

equipped with homogeneous boundary and final-time conditions, measures the distance to the boundary. Due to (6) the eikonal equation belongs to the Hamilton-Jacobi-Bellman family. We consider the equation on a domain (Fig. 4, top left) whose irregular shape leads to complicated curves on which v is not differentiable. The height of the domain is equal to one.

Figure 4 compares locally-adapted choices of the artificial diffusion parameter with global choices, and illustrates the effect of mesh refinement on the quality of the solution. To compare the quality of the numerical solutions we compare their L^∞-norms—recalling that excessive numerical dissipation leads to a smearing out of extrema. The coarse grid solution, with 2,858 internal nodes and with a global diffusion parameter of 0.05, only reaches a height of 9.56×10^{-2}. In contrast, a computation on the same mesh with a local diffusion parameter, with mean value

0.01 and standard deviation 0.005 but maximal value 0.05, leads to a height of 0.138. With one (two) steps of uniform refinement the number of elements increases by a factor of 4 (of 16) and the artificial diffusion decreases to an average value of 0.004 (of 0.002), giving the improved value of 0.147 (of 0.151) for the L^∞-norm in the lower left (right) part of Fig. 4. The L^∞-norm of a reference solution on a very fine mesh is 0.153.

6 A Second-Order Fully Nonlinear Equation

The final example, on the same domain as in the previous section, is concerned with the fully nonlinear case of second-order. We examine the equation

$$-v_t + \sup_{\alpha\in[\alpha_0,\alpha_1]} \{-\alpha\Delta v\} + |\nabla v| = -v_t + \sup_{(\alpha,\beta)\in[\alpha_0,\alpha_1]\times\partial B(0,1)} \{-\alpha\Delta v + \beta\cdot\nabla v\} = f,$$

where $B(0,1)$ is the unit ball in $\mathbb{R}^2$, $T = 0.009$, $\alpha_0 = 0.045$, $\alpha_1 = 0.09$ and

$$f(x,y) = 529\left(\sin\big(g(x,y)\big) + \frac{1}{2}\sin\big(2\,g(x,y)\big) + \frac{4}{10}\sin\big(8\,g(x,y)\big)\right)^2,$$

$$g(x,y) = \pi^2(x-0.63)(y-0.26)/0.07.$$

The boundary and final-time conditions are homogeneous. As before, the advection term is discretised explicitly, with the locally minimal diffusion needed for monotonicity. The possibly remaining (fully nonlinear) diffusion is placed in the implicit term.

In this example the control α of the second-order term is maximised independently of the first-order control β; in the sense that the optimal α may be determined without knowledge of β. Furthermore, as Δv takes locally either a positive or negative value only the controls α_0 and α_1 are ever active in the HJB equation. This is an example of the Bang-bang principle. It is also reflected in the left plot in Fig. 5, where the value of α is plotted. Black colouring signifies α_0 maximises the operator, whereas white colouring corresponds to α_1. Observe that no intermediate grey values can appear. The plot of the control α mimics some of the features of the value function, which is plotted in the same figure on the right—in part because the Laplacian contains information about the curvature of the solution.

At each time-step of the method, a semi-smooth Newton method [4] was used to solve the nonlinear discrete equation in (4), where each iteration of the algorithm involves solving a linear system. To study the performance of the algorithm, the HJB equation was solved on a sequence of successively refined meshes, with a constant set of tolerances. The linear systems were solved to a tolerance of 10^{-10} by GMRES. The stopping criterion for Newton's method was a relative residual tolerance of 5×10^{-8} in the maximum norm and a convergence of the iterations requirement of

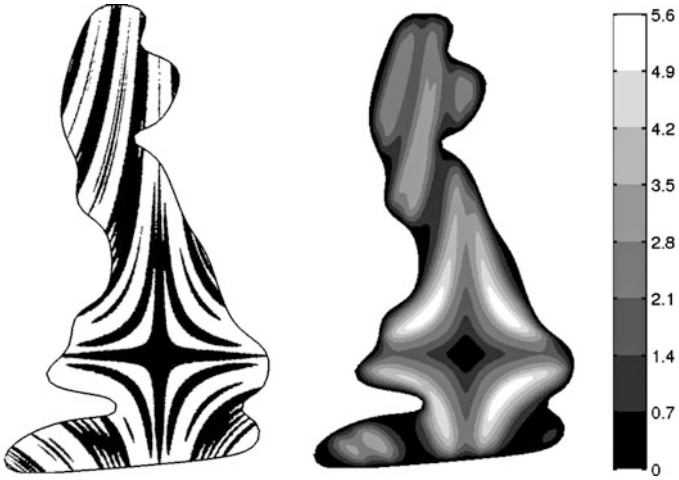


Fig. 5 The *left plot* shows the contours of the control α maximising the nonlinear second order term of the equation, the *right plot* shows the value function v. In the *left plot*, the *black regions* correspond to $\alpha = \alpha_0$, whilst the *white regions* correspond to $\alpha = \alpha_1$

5×10^{-9} in the maximum norm. The sizes of the linear systems for the sequence of meshes were 674, 2,858, 11,759, 47,693 and 192,089. The respective average number of Newton iterations for a single time step were 3, 3.67, 4.04, 4.22 and 4.86, with respective standard deviations 0, 0.48, 0.19, 0.42 and 0.36. This demonstrates a weak dependence of the number of iterations needed for an individual time step on the system size, thus showing that the total number of linear systems to be solved for a complete computation depends principally on the number of time-steps.

References

1. G. Barles and P.E. Souganidis, *Convergence of approximation schemes for fully nonlinear second order equations*, J. Asymptotic Analysis, 4:271–283, 1991.
2. X. Feng and R. Glowinsky and M. Neilan, *Recent developments in numerical methods for fully nonlinear second order partial differential equations*, submitted to SIAM Review.
3. X. Feng and M. Neilan, *The vanishing moment method for fully nonlinear second order partial differential equations: formulation, theory, and numerical analysis*, arXiv:1109.1183, 2011.
4. M. Jensen and I. Smears, *On the convergence of finite element methods for Hamilton-Jacobi-Bellman equations*, arXiv:1111.5423, 2011.
5. H.J. Kushner and P. Dupuis, *Numerical methods for stochastic control problems in continuous time*, Applications of Mathematics 24, Springer-Verlag, 2001.

Adaptive Computation of Parameters in Stabilized Methods for Convection-Diffusion Problems

V. John and P. Knobloch

Abstract Stabilized finite element methods for convection-dominated problems contain parameters whose optimal choice is usually not known. This paper presents techniques for computing stabilization parameters in an adaptive way by minimizing a target functional characterizing the quality of the approximate solution. This leads to a constrained nonlinear optimization problem. Numerical results obtained for various target functionals are presented. They demonstrate that a posteriori optimization of parameters can significantly improve the quality of solutions obtained using stabilized methods.

1 Introduction

This paper is devoted to the numerical solution of a steady scalar convection-diffusion equation

$$-\varepsilon \Delta u + \mathbf{b} \cdot \nabla u + c\, u = f \quad \text{in } \Omega\,, \qquad u = u_b \quad \text{on } \partial\Omega \tag{1}$$

by means of the finite element method. In (1), $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, is a bounded domain with a polygonal (resp. polyhedral) Lipschitz-continuous boundary $\partial\Omega$,

V. John
Weierstrass Institute for Applied Analysis and Stochastics (WIAS), Mohrenstr. 39, 10117 Berlin, Germany

Department of Mathematics and Computer Science, Free University of Berlin, Arnimallee 6, 14195 Berlin, Germany
e-mail: volker.john@wias-berlin.de

P. Knobloch (✉)
Department of Numerical Mathematics, Faculty of Mathematics and Physics, Charles University, Sokolovská 83, 186 75 Praha 8, Czech Republic
e-mail: knobloch@karlin.mff.cuni.cz

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_30,

$\varepsilon > 0$ is constant, $\mathbf{b} \in W^{1,\infty}(\Omega)^d$, $c \in L^\infty(\Omega)$, $f \in L^2(\Omega)$, and $u_b \in H^{1/2}(\partial\Omega)$. The Dirichlet boundary condition is used for the sake of simplicity only. In the numerical computations presented in this paper also more general boundary conditions were used.

Problem (1) is a simple model problem for convection-diffusion effects appearing in many more complicated applications. Therefore, it is important to be able to solve this problem numerically in a satisfactory way. However, this is by no means easy if convection dominates diffusion, i.e., $\varepsilon \ll |\mathbf{b}|$, since then the solution of (1) contains so-called layers, which are narrow regions where the solution changes abruptly. It is well known that the standard Galerkin finite element method provides approximate solutions that are globally polluted by spurious oscillations unless the computational mesh is sufficiently fine, i.e., $\varepsilon \gtrsim |\mathbf{b}|\, h$ where h is the mesh parameter.

To suppress the spurious oscillations, there are basically two options. Either one can use a layer-adapted mesh (e.g., a piecewise uniform mesh or a mesh obtained by an anisotropic adaptive refinement strategy) or one can consider a relatively coarse mesh and employ a modification of the standard discretization. There are various modifications that can be found in the literature: special discretizations of the convective term (upwinding), introduction of additional terms (stabilization) or manipulations at the algebraic level (e.g., FEMTVD schemes). In this paper, we shall be interested in stabilization techniques applied on relatively coarse meshes.

A common feature of stabilized finite element methods is that they contain parameters whose values significantly influence the quality of the approximate solution but whose optimal choice is usually not known. The aim of the present paper is to describe techniques that make it possible to compute stabilization parameters in an adaptive way by minimizing a functional characterizing the quality of the approximate solution. This leads to a constrained nonlinear optimization problem. The paper is a continuation of our previous work published in [3] where basic ideas of the optimization of stabilization parameters were presented.

The plan of the paper is as follows. In the next two sections we discuss linear and nonlinear stabilization approaches for finite element discretizations of (1). Then, in Sect. 4, we describe our approach of parameter optimization and explain how the Fréchet derivative of the target functional can be computed in an efficient way. Finally, in Sect. 5, we construct several target functionals and illustrate their properties by means of numerical results.

2 Linear Stabilized Methods

Let W_h be a finite element space approximating the space $H^1(\Omega)$ and set $V_h := W_h \cap H_0^1(\Omega)$. Let $u_{bh} \in W_h$ be a function whose trace approximates the function u_b. The simplest finite element discretization of (1) is the Galerkin method that reads: Find $u_h \in W_h$ such that $u_h = u_{bh}$ on $\partial\Omega$ and

$$a(u_h, v_h) = (f, v_h) \qquad \forall\ v_h \in V_h\,,$$

where $a(u,v) = \varepsilon(\nabla u, \nabla v) + (\mathbf{b}\cdot\nabla u, v) + (c\,u, v)$ and $(\cdot,\cdot)$ denotes the inner product in $L^2(\Omega)$ or $L^2(\Omega)^d$. As we mentioned in the introduction, the Galerkin discretization is not appropriate if convection dominates diffusion and, as a remedy, a stabilization of the Galerkin method will be considered.

A stabilized finite element method for the numerical solution of (1) can be obtained from the Galerkin method by adding a stabilization term. We shall consider methods that read: Find $u_h \in W_h$ such that $u_h = u_{bh}$ on $\partial\Omega$ and

$$a(u_h, v_h) + \sum_{K\in\mathscr{T}_h} \tau_K\, s_K(u_h, v_h) = (f, v_h) \qquad \forall\ v_h \in V_h\,.$$

Here $\mathscr{T}_h$ is the triangulation used for constructing the finite element space W_h, τ_K is a nonnegative stabilization parameter, and s_K is a local form whose arguments are functions defined on the set $K \in \mathscr{T}_h$. The form s_K is always linear in the second argument and, if $f = 0$, it is also linear in the first argument. There are examples of s_K which are bilinear for any f. The parameter τ_K determines the artificial diffusion added by the stabilization term and it should be not 'too small' to remove oscillations but also not 'too large' to avoid excessive smearing. Consequently, it is very difficult to find appropriate values of τ_K a priori.

One of the most popular finite element approaches for convection-dominated problems is the SUPG method for which

$$s_K(u,v) = (\mathscr{L}_h\, u - f, \mathbf{b}\cdot\nabla v)_K$$

with the differential operator $\mathscr{L}_h = -\Delta_h + \mathbf{b}\cdot\nabla + c$ where the subscript h indicates that the Laplace operator is applied elementwise. The stabilization parameter is often defined by

$$\tau_K = \frac{h_K}{2\,|\mathbf{b}|}\left(\coth \mathrm{Pe}_K - \frac{1}{\mathrm{Pe}_K}\right) \quad \text{with} \quad \mathrm{Pe}_K = \frac{|\mathbf{b}|\,h_K}{2\,\varepsilon}\,, \tag{2}$$

where h_K is the diameter of K in the direction of $\mathbf{b}$.

3 Nonlinear Stabilized Methods

Since solutions of linear stabilized methods usually possess spurious oscillations in layer regions, the so-called SOLD (spurious oscillations at layers diminishing) methods have been developed. These methods add an additional stabilization term to the left-hand side of a linear stabilized method. Typical examples of this term are $(\widetilde{\varepsilon}\nabla u_h, \nabla v_h)$ adding isotropic artificial diffusion and $(\widetilde{\varepsilon}\,P\,\nabla u_h, P\,\nabla v_h)$ with the orthogonal projection P onto the plane orthogonal to $\mathbf{b}$, adding crosswind artificial diffusion. The parameter $\widetilde{\varepsilon}$ usually depends on the unknown approximate solution u_h and hence the resulting method is nonlinear.

In the literature, many proposals for the parameter $\widetilde{\varepsilon}$ can be found and we refer to [1, 2] for a review and computational comparison. One of the most successful formulas is

$$\widetilde{\varepsilon}|_K = \eta \frac{\operatorname{diam}(K)\,|\mathscr{L}_h u_h - f|}{2\,|\nabla u_h|} \qquad \forall\ K \in \mathscr{T}_h\,,$$

where η is a user-chosen parameter. From now on, the notion 'SOLD method' will mean that the crosswind diffusion term $(\widetilde{\varepsilon} P \nabla u_h, P \nabla v_h)$ together with this choice of $\widetilde{\varepsilon}$ is used. In the framework of parameter optimization, the parameter η will be considered piecewise constant. If an optimization of η is not considered, we set $\eta = 0.7$.

4 A Posteriori Optimization of Stabilization Parameters

In this section, we describe basic ideas of our approach to a posteriori optimization of stabilization parameters. For clarity of the presentation, we shall restrict ourselves to $u_b = 0$.

Let us write a linear or nonlinear stabilized method in the abstract form:

Given a stabilization parameter $y_h \in Y_h$, find $u_h \in V_h$ such that $R_h(u_h, y_h) = 0$.

Here, Y_h is a finite-dimensional space of functions on Ω and the operator R_h maps the space $V_h \times Y_h$ into the dual space V_h'. For example, for the SUPG method introduced in Sect. 2, we have

$$\langle R_h(u_h, y_h), v_h\rangle = a(u_h, v_h) + (\mathscr{L}_h u_h - f, y_h\,\mathbf{b}\cdot\nabla v_h) - (f, v_h)$$

and Y_h can be the space of piecewise constant functions on Ω. To emphasize that the approximate solution u_h depends on the choice of the stabilization parameter $y_h \in Y_h$, we shall write $u_h(y_h)$ instead of u_h in the following.

We introduce a functional $I_h : V_h \to \mathbb{R}$ such that $I_h(u_h(y_h))$ represents a measure of the error or the quality of $u_h(y_h)$. We assume that the solution $u_h(y_h)$ improves if the functional $\Phi_h(y_h) := I_h(u_h(y_h))$ decreases. Thus, our aim is to find $y_h \in Y_h$ such that $\Phi_h(y_h)$ is 'small'. This is a constrained nonlinear optimization problem since y_h has to be nonnegative and smaller than some upper bound. For example, for the SUPG method,

$$0 \le y_h|_K \le 10\,\tau_K \qquad \forall\ K \in \mathscr{T}_h\,, \tag{3}$$

where τ_K is defined by (2). The factor 10 can be changed to another value but numerical experiments indicate that the factor should not differ too much from 10.

Common minimization algorithms require at least the knowledge of the derivative of the function which should be minimized. Thus, we have to compute the Fréchet derivative of the functional Φ_h. Using the chain rule, we obtain

$$D\Phi_h(y_h) = DI_h(u_h(y_h))Du_h(y_h)\,.$$

However, it is not efficient to compute $D\Phi_h(y_h)$ using this formula since it requires the solution of $\dim Y_h$ linear problems of the size of the original discrete problem. Therefore, we first define the adjoint problem: Find $\psi_h(y_h) \in V_h$ such that

$$(\partial_u R_h)'(u_h(y_h), y_h)\, \psi_h(y_h) = DI_h(u_h(y_h))\,,$$

where $\langle(\partial_u R_h)'(w_h, y_h)v_h, \tilde{v}_h\rangle = \langle(\partial_u R_h)(w_h, y_h)\tilde{v}_h, v_h\rangle \quad \forall\ v_h, \tilde{v}_h, w_h \in V_h$, $y_h \in Y_h$. Since $R_h(u_h(y_h), y_h) = 0$, we have $\partial_u R_h(u_h(y_h), y_h)Du_h(y_h) + \partial_y R_h(u_h(y_h), y_h) = 0$. Thus, combining the above relations, we deduce that

$$D\Phi_h(y_h) = -(\partial_y R_h)'(u_h(y_h), y_h)\psi_h(y_h)\,,$$

where $\langle(\partial_y R_h)'(w_h, y_h)v_h, \tilde{y}_h\rangle = \langle(\partial_y R_h)(w_h, y_h)\tilde{y}_h, v_h\rangle \quad \forall\ v_h, w_h \in V_h, y_h, \tilde{y}_h \in Y_h$. Note that, for the SUPG method, the function $\psi_h(y_h)$ solves

$$a(v_h, \psi_h(y_h)) + (\mathscr{L}_h\, v_h, y_h\, \mathbf{b}\cdot\nabla\psi_h(y_h)) = \langle DI_h(u_h(y_h)), v_h\rangle \quad \forall\ v_h \in V_h$$

and the Fréchet derivative of Φ_h is given by

$$\langle D\Phi_h(y_h), \tilde{y}_h\rangle = -(\mathscr{L}_h\, u_h(y_h) - f, \tilde{y}_h\, \mathbf{b}\cdot\nabla\psi_h(y_h))\,.$$

5 Choice of the Functional I_h

In this section, we propose various choices of the functional I_h introduced in the previous section and present numerical results illustrating the properties of these functionals.

All numerical results were computed for $\Omega = (0,1)^2$ and, in all cases, we considered a triangulation $\mathscr{T}_h$ of Ω constructed by dividing Ω into 32×32 equal squares and each square into two triangles by drawing a diagonal from bottom right to top left. The space W_h consisted of continuous piecewise linear functions. The functional Φ_h was minimized using the BFGS method [4]. The SUPG parameter was initialized by (2) and the SOLD parameter by 0. The SUPG parameter satisfied the constraints (3) and the SOLD parameter was required to be in the interval $[0, 1]$.

In each iteration of the BFGS method, one has to solve once the adjoint problem and several times the discrete problem for various values of the stabilization parameter. Consequently, the cost of the computation of an optimized SUPG stabilization parameter is significantly higher than the computation of the SUPG solution for a prescribed stabilization parameter. Comparing the cost of the optimization with the cost of the solution of a nonlinear SOLD method, the difference is not so large. We believe that the higher computational cost of the parameter optimization is justified by the quality of the resulting approximate solution, cf. the examples in this section.

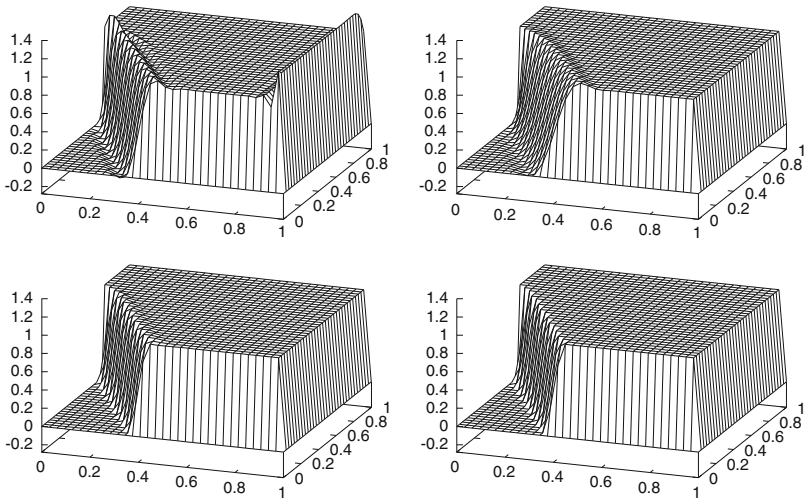


Fig. 1 Example 1: SUPG standard (*top left*), SUPG optimized using I_h^{res} (*top right*), SUPG optimized using $I_h^{\text{res}} + \alpha\, I_h^{\text{cross}}$ (*bottom left*), SOLD optimized using $I_h^{\text{res}} + \alpha\, I_h^{\text{cross}}$ (*bottom right*)

We denote by $\Gamma^+ = \overline{\{\mathbf{x} \in \partial\Omega\,;\ (\mathbf{b}\cdot\mathbf{n})(\mathbf{x}) > 0\}}$, $\Gamma^0 = \overline{\{\mathbf{x} \in \partial\Omega\,;\ (\mathbf{b}\cdot\mathbf{n})(\mathbf{x}) = 0\}}$ the outflow and characteristic boundaries of Ω, respectively. Furthermore, we set

$$G_h = \bigcup_{K \in \mathscr{G}_h} \overline{K} \qquad \text{with} \qquad \mathscr{G}_h = \{K \in \mathscr{T}_h\,;\ \overline{K} \cap \Gamma^+ \neq \emptyset \ \text{ or } \ \overline{K} \cap \Gamma^0 \neq \emptyset\}\,.$$

Note that G_h represents a strip along Γ^+ and Γ^0 made up of elements of $\mathscr{T}_h$ having at least one vertex on these parts of the boundary. A functional characterizing the quality of an approximate solution u_h of (1) can be now defined by

$$I_h^{\text{res}}(u_h) = \|\mathscr{L}_h\, u_h - f\,\|^2_{0,\Omega\setminus G_h}\,.$$

We exclude the strip G_h since even a nodally exact solution has a large error in G_h. Let us apply the functional I_h^{res} to the numerical solution of the following example.

Example 1 (Solution with an interior layer and two exponential boundary layers). We consider the convection-diffusion equation (1) with $\Omega = (0,1)^2$, $\varepsilon = 10^{-8}$, $\mathbf{b} = (\cos(-\pi/3), \sin(-\pi/3))^T$, $c = f = 0$, $u_b(x,y) = 0$ for $x = 1$ or $y \leq 0.7$, and $u_b(x,y) = 1$ else. The function u_b could also be replaced by a function from $H^{1/2}(\partial\Omega)$ leading to the same numerical results as presented in this paper.

Figure 1 (top left) shows the SUPG solution computed with the stabilization parameter τ_K given by (2). If we optimize the stabilization parameter using the functional I_h^{res}, the spurious oscillations along the exponential boundary layer are

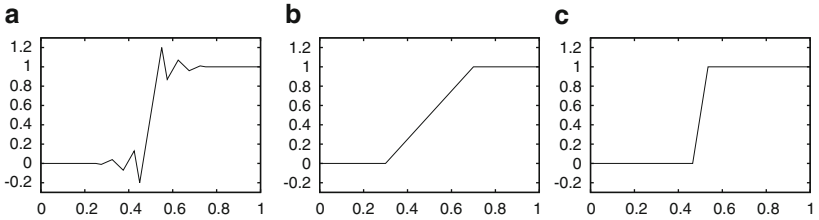


Fig. 2 Idealized cuts through approximate solutions across an interior layer

removed but those along the interior layer are not suppressed sufficiently. Moreover, the interior layer is smeared, see Fig. 1 (top right).

If we observe a cut through the solution in Fig. 1, top left, across the interior layer, we shall see a curve like in Fig. 2a. We would like to compute a solution without spurious oscillations corresponding to Fig. 2b or c. A candidate for a functional which prefers a solution without spurious oscillations is $\int_0^1 |u'|^p \mathrm{d}x$, where u represents the functions in Fig. 2. Denoting by d the width of the layer in Fig. 2b or c, the integral equals d^{1-p}. Since we prefer the curve c, we have to use $p < 1$. Thus, we may consider the functional

$$I_h^{\text{cross}}(u_h) = \int_{\Omega \setminus G_h} \sqrt{|\mathbf{b}^\perp \cdot \nabla u_h|}\, \mathrm{d}\mathbf{x}\,,$$

where $\mathbf{b}^\perp$ is a unit vector orthogonal to $\mathbf{b}$. In our implementation, the square root is regularized near 0, see [3] for details. If we now optimize the SUPG stabilization parameter using a combination of I_h^{res} and I_h^{cross}, the solution improves considerably, see Fig. 1 (bottom left). Finally, if we perform the optimization with the same functional but for the SOLD method, we obtain a solution without any visible spurious oscillations and with steep layers, see Fig. 1 (bottom right).

Example 2 (Solution with one exponential and two parabolic boundary layers). We consider the convection-diffusion equation (1) with $\Omega = (0,1)^2$, $\varepsilon = 10^{-8}$, $\mathbf{b} = (1,0)^T$, $c = 0$, $f = 1$, and $u_b = 0$.

For this example, a comparison of the SUPG solution without parameter optimization and an optimized SOLD solution is given in Fig. 3. It can be observed, that the parameter optimization leads to an almost nodally exact solution.

Example 3 (Solution with two interior layers). We consider the convection-diffusion equation (1) with $\Omega = (0,1)^2$, $\varepsilon = 10^{-8}$, $\mathbf{b}(x,y) = (-y,x)^T$, and $c = f = 0$. On $\Gamma^N := \{0\} \times (0,1)$, we prescribe a homogeneous Neumann boundary condition whereas the Dirichlet boundary condition is considered only on $\Gamma^D := \partial\Omega \setminus \overline{\Gamma^N}$ with $u_b(x,y) = 1$ for $(x,y) \in (1/3, 2/3) \times \{0\}$ and $u_b(x,y) = 0$ else on Γ^D.

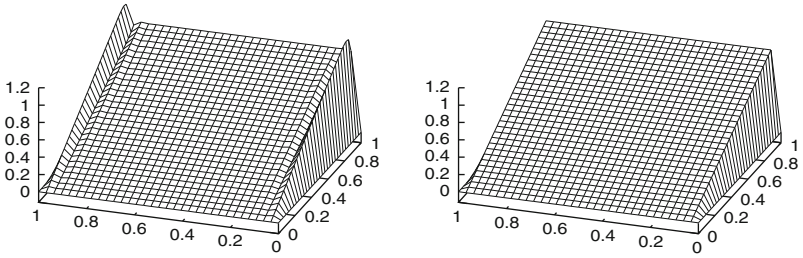


Fig. 3 Example 2: SUPG standard (*left*), SOLD optimized using I_h^{res} (*right*)

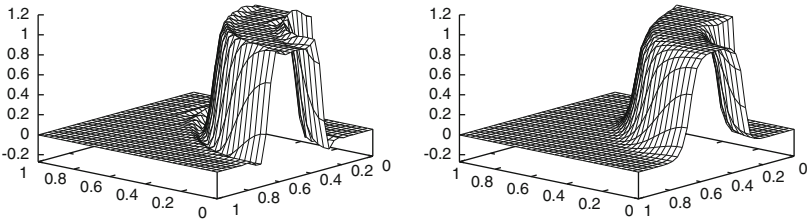


Fig. 4 Example 3: SUPG standard (*left*), SOLD standard (*right*)

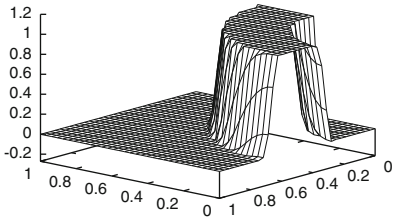


Fig. 5 Example 3: SOLD optimized using I_h^{cross}

Figure 4 shows results for this example obtained without parameter optimization. We see that the SOLD method suppresses the oscillations present in the SUPG solution but leads to a slight smearing of the layers. The quality of the SOLD solution obtained using parameter optimization is much better, see Fig. 5.

Acknowledgements The work of P. Knobloch is a part of the research project MSM 0021620839 financed by MSMT and it was partly supported by the Grant Agency of the Czech Republic under the grant No. 201/08/0012.

References

1. V. John and P. Knobloch. On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations: Part I – A review. *Comput. Methods Appl. Mech. Engrg.*, 196:2197–2215, 2007.
2. V. John and P. Knobloch. On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations: Part II – Analysis for P_1 and Q_1 finite elements. *Comput. Methods Appl. Mech. Engrg.*, 197:1997–2014, 2008.
3. V. John, P. Knobloch, and S. B. Savescu. A posteriori optimization of parameters in stabilized methods for convection–diffusion problems – Part I. *Comput. Methods Appl. Mech. Engrg.*, 200:2916–2929, 2011.
4. J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer, 2nd edition, 2006.

The Numerical Study of Singularly Perturbed Differential-Difference Turning Point Problems: Twin Boundary Layers

P. Rai and K.K. Sharma

Abstract A boundary value problem for singularly perturbed differential-difference equation with turning point is considered. Some a priori estimates are obtained on the solution and its derivatives. In general, to tackle such type of problems one encounters three difficulties: (i) due to presence of the turning point, (ii) due to presence of terms containing shifts and (iii) due to presence of the singular perturbation parameter. Due to presence of the singular perturbation parameter the classical numerical methods fail to give reliable numerical results and do not converge uniformly with respect to the singular perturbation parameter. In this paper a parameter uniform finite difference scheme is constructed to solve the boundary-value problem. A parameter uniform error estimate for the numerical scheme so constructed is established. Numerical experiments are carried out to demonstrate the efficiency of the numerical scheme and support the theoretical estimates.

1 Introduction and Problem Formulation

It is a well-established principle to model the evolution of physical, biological and economic system using ordinary or partial differential equations in which the response of the system depends purely on the current state of the system but there are many cases in which the response of the system depend upon the past history of the system. Dynamical systems which respond in this way are called delay differential equations (DDEs). Furthermore, in applications the system can be perturbed by noise, be intrinsically random or in which certain parameters in the model are unknown. In these cases, it is more appropriate to model the dynamics

P. Rai (✉) · K.K. Sharma
Department of Mathematics (Center for Advance Study in Mathematics), Panjab University, Chandigarh, 160014, India
e-mail: pratimarai5@gmail.com; kapilks@pu.ac.in

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_31,

of the system using stochastic delay differential equations (SDDEs). The work of the delay differential equations group involves the study of both DDEs and SDDEs, concentrating in particular on their long time behavior. There are some cases where both type of shifts, i.e., delay as well as advance arguments are present and such type of equations are called differential-difference equations.

Differential-difference equations govern a variety of physical processes, for instance, hydrodynamics of liquid helium [1], thermoelasticity [2], study of variational problems in control theory [4], diffusion in polymers [5], study of bistable devices [7], evolutionary biology [8], micro scale heat transfer, description of human pupil light reflex [6], a variety of models of physiological processes or diseases [8, 20]. They are also satisfied by the moments of the time of first exit of temporally homogeneous Markov processes [9] governing such phenomena as the time between impulses of a nerve cell and the persistence time of populations with large random fluctuations.

Singular perturbation problems are the differential equations where the highest order derivative term is multiplied by a small parameter ε which can take arbitrary value between $(0, 1]$. Turning point is a point of the domain where the coefficient of the convection term vanishes. The solution of such type of differential equations exhibits boundary layer(s) or interior layer(s) behavior depending upon the nature of the coefficient of the convection and the reaction term. In this paper we are interested in studying the case where twin boundary layer exist in the solution of the problem due to presence of the turning point.

Study of singularly perturbed differential-difference equation was initiated by Lange and Miura. They gave a series of paper [10–12] where they did asymptotic study of such type of problems and discussed the case of small as well as large delay. Kadalbajoo and Sharma [14–19] initiated the numerical study of singularly perturbed differential-difference equations and studied the effect of small and large delay as well as advance on the layer behavior of the solution. They considered the case where the coefficient of the convection term has same sign throughout the domain but the case of turning point is still unexplored and there is a lot to study in this case. Rai and Sharma [21,22] investigated the singularly perturbed turning point problems with interior layer and studied both the cases, i.e., when shift are $o(\varepsilon)$ [21] as well as the case when they are $O(\varepsilon)$ [22]. In this paper, we initiate the numerical study of singularly perturbed differential-difference equations with turning points and exhibiting twin boundary layers.

We consider the following singularly perturbed differential-difference equation having isolated turning point at $x = 0$

$$\begin{aligned} &\varepsilon y''(x) + a(x)y'(x) - b(x)y(x) + c(x)y(x-\delta) + d(x)y(x+\eta) = f(x), \quad x \in \Omega = (-1, 1) \\ &y(x) = \varphi(x), \ -1-\delta \le x \le -1, \quad y(x) = \gamma(x), \ 1 \le x \le 1+\eta \end{aligned} \tag{1}$$

where $\bar{\Omega} = [-1, 1]$, $\bar{\Omega}_1 = (-1, -1+\delta)$ $\bar{\Omega}_2 = (-1+\delta, 1-\eta)$ $\bar{\Omega}_3 = (1-\eta, 1)$ $\delta, \eta = o(\varepsilon)$, $0 < \varepsilon \ll 1$, $a(x)$, $b(x)$, $f(x)$, $\varphi(x)$ are sufficiently smooth functions. Further, it is also assumed that

$$a(0) = 0,\ a'(0) < 0 \tag{2}$$

$$b(x) - c(x) - d(x) \geq k_0 > 0,\ c(x) \geq 2M_1 > 0,\ d(x) \geq 2M_2 > 0 \quad \forall x \in [-1, 1] \tag{3}$$

$$|a'(x)| > |a'(0)|/2,\ \forall x \in [-1, 1]. \tag{4}$$

2 A Priori Estimates

Use of Taylor's series expansion to approximate the shift arguments gives us

$$y(x - \delta) \approx y(x) - \delta y'(x) + \frac{\delta^2}{2} y''(x) \tag{5}$$

$$y(x + \eta) \approx y(x) + \eta y'(x) + \frac{\eta^2}{2} y''(x). \tag{6}$$

Substituting the above approximation in (1) results into

$$\begin{aligned} &L_\varepsilon^1 y(x) \equiv C_\varepsilon(x) y''(x) + A(x) y'(x) - B(x) y(x) = f(x) \\ &y(-1) = \varphi(-1), \quad y(1) = \gamma(1), \end{aligned} \tag{7}$$

where $C_\varepsilon(x) = (\varepsilon + \frac{\delta^2}{2} c(x) + \frac{\eta^2}{2} d(x)) > 0$, $A(x) = a(x) - \delta c(x) + \eta d(x)$, $B(x) = b(x) - c(x) - d(x)$.

The solution of the problem (7) is an approximation to the solution of the problem (1). The problem (7) satisfy following minimum principle

Lemma 1. *Let $\psi(x)$ be a smooth function satisfying $\psi(-1) \geq 0$, $\psi(1) \geq 0$ and $L_\varepsilon^1 \psi(x) \leq 0\ \forall x \in \Omega$. Then $\psi(x) \geq 0$, $\forall\ x \in \bar{\Omega}$.*

Using above minimum principle it is easy to prove that

Lemma 2. *The solution $y(x)$ of the problem (7) satisfies*

$$||y|| \leq ||f||/k_0 + max\ (|\varphi(-1)|,\ |\gamma(1)|). \tag{8}$$

Now, to derive bounds on the derivatives of the solution of the problem (7) we have following results.

Theorem 1. *If $y(x)$ is solution of the problem (7) and $|a(x)| \geq a_0 > 0$, $\forall x \in \Omega_1 \cup \Omega_3$ then, there exist a positive constant C such that for $a(x) < 0$ we have*

$$|y^{(k)}(x)| \leq C \left(1 + C_\varepsilon^{-k} exp(-a_0(1 - x)/C_\varepsilon)\right),\ x \in \Omega_3,\ k = 1, \ldots 3$$

and for $a(x) > 0$ we have

$$|y^{(k)}(x)| \leq C \left(1 + C_\varepsilon^{-k} exp(-a_0(x + 1)/C_\varepsilon)\right),\ x \in \Omega_1,\ k = 1, \ldots 3.$$

Theorem 2. *Assume (2)–(4) and* $S_2 = \{||a||_2,\ ||f||_2,\ ||b||_2,\ k_0,\ |\varphi(-1)|,\ |\gamma(1)|\}$ *then, there exist a positive constant* C *depending upon* S_2 *such that*

$$|y^{(k)}(x)| \leq C,\ x \in \Omega_2,\ k = 1, 2, 3.$$

Theorem 3. *The solution* $y(x)$ *of the Problem (7) admits the decomposition*

$$y(x) = v_\varepsilon(x) + w_\varepsilon(x)$$

where the regular component $v_\varepsilon(x)$ satisfies

$$|v_\varepsilon^{(k)}(x)| \leq C\left(1 + C_\varepsilon^{(2-k)} e(x, a_0)\right),\ x \in \bar{\Omega},\ k = 1, 2, 3$$

and the singular component $w_\varepsilon(x)$ satisfies

$$|w_\varepsilon^{(k)}(x)| \leq M C_\varepsilon^{-k} e(x, a_0),\ x \in \bar{\Omega},\ k = 1, 2, 3$$

where $e(x, a_0) = (exp(-a_0(1-x)/C_\varepsilon) + exp(-a_0(x+1)/C_\varepsilon))$.

3 Numerical Discretization

In this subsection we discuss an upwind finite difference scheme composed of classical upwind scheme on a piecewise uniform Shishkin mesh [13] $\bar{\Omega}^N$, condensing at the boundaries $x = -1,\ 1$ for the boundary value problem (7). The fitted piecewise uniform mesh $\bar{\Omega}^N$ is constructed by partitioning the interval $[-1, 1]$ into three subintervals namely, $\bar{\Omega}_1^N = [-1, -1+\tau]$, $\bar{\Omega}_2^N = [-1+\tau, 1-\tau]$, $\bar{\Omega}_3^N = [1-\tau, 1]$ where the transition parameter τ is given by

$$\tau = \min(1/2, KC_\varepsilon^* \ln N), \quad \text{where } K = \frac{1}{\min(a_0, k_0)},\ C_\varepsilon^* = \varepsilon + \delta^2 M_1 + \eta^2 M_2.$$

We define the difference operator for the problem (7) by

$$\begin{aligned} &L_{\varepsilon,N}^1 Y_i = f_i \quad \text{for } i = 1, 2 \ldots, N-1 \\ &Y_0 = \varphi(-1), \quad Y_N = \gamma(1) \end{aligned} \tag{9}$$

where

$$L_{\varepsilon,N}^1 Y_i = C_\varepsilon(x_i) D^+ D^- Y_i + A(x_i) D^* Y_i - B(x_i) Y_i \tag{10}$$

and $T_i = T(x_i)$, $D^+D^-T_i = \frac{2(D^+T_i - D^-T_i)}{x_{i+1} - x_{i-1}}$, $D^+T_i = \frac{T_{i+1} - T_i}{x_{i+1} - x_i}$, $D^-T_i = \frac{T_i - T_{i-1}}{x_i - x_{i-1}}$,

$$D^*T_i = \begin{cases} D^+T_i, & a(x_i) > 0 \\ D^-T_i, & a(x_i) < 0 \end{cases}, \quad h_i = \begin{cases} 4\tau/N,\ 0 \leq i \leq N/4 \\ 4(1-\tau)/N, \quad N/4 + 1 \leq i \leq 3N/4 \\ 4\tau/N, \quad 3N/4 + 1 \leq i \leq N \end{cases}$$

and

$$x_i = \begin{cases} -1 + ih_i & \text{for } 0 \le i \le N/4 \\ \tau - 1 + (i - N/4)h_i & \text{for } N/4 + 1 \le i \le 3N/4 \\ 1 - \tau + (i - 3N/4)h_i & \text{for } 3N/4 + 1 \le i \le N. \end{cases}$$

Lemma 3. *Suppose* $\Psi_0 \ge 0$ *and* $\Psi_N \ge 0$. *Then* $L^1_{\varepsilon,N}\Psi_i \le 0,\ i = 1(1)N - 1$ *implies that* $\Psi_i \ge 0,\ i = 0(1)N$.

Lemma 4. *Let* U_i *be any mesh function such that* $U_0 = U_N = 0$. *Then, for all* $i,\ 0 \le i \le N$

$$|U_i| \le \frac{1}{k_0} max_{1 \le j \le N-1} |L^1_{\varepsilon,N} U_j|.$$

Theorem 4. *The solution* Y^N *of the discrete boundary value problem* (9) *and the solution* $y(x)$ *of the continuous boundary value problem* (7) *satisfies the error estimate*

$$sup_{0<\varepsilon\le 1} ||Y^N(x_i) - y(x_i)|| \le CN^{-1}(ln\ N)^2 \quad x_i \in \bar{\Omega}^N.$$

4 Numerical Results

In this section we present some numerical experiments to show efficiency of the discussed algorithm. Since exact solution is not known double mesh principle is used for the calculation of maximum pointwise error [3].

Example 1.

$$\varepsilon y''(x) + 2(2x-1)y'(x-\delta) - 3y(x) + y(x-\delta) + 0.5y(x+\eta) = 0, \quad x \in (0,1)$$

under interval and boundary conditions

$$y(x) = 1, \quad -\delta \le x \le 0 \text{ and } y(x) = 1 \quad 1 \le x \le 1 + \eta.$$

5 Discussion

Objective of the present study is to investigate singularly perturbed differential-difference turning point problems exhibiting twin boundary layers. We use Taylor's series to approximate the retarded arguments and standard finite difference scheme is used to approximate the resulting differential equation on appropriately constructed Shishkin mesh. Theoretical error bounds are established and uniform

Table 1 Maximum pointwise errors ($E_{N,\epsilon}$) and experimental order of convergence (EOC), when applied to Example 1 for various values of ϵ and N and $\delta = \varepsilon/10,\ \eta = \varepsilon/5$

ϵ	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$	$N = 512$	N =1,024
2^{-2}	2.67295E − 02	1.46677E − 02	7.71286E − 03	3.96144E − 03	2.00809E − 03	1.01110E − 03	5.07329E − 04
	0.87	0.93	0.96	0.98	0.99	0.99	1.0
2^{-6}	8.24259E − 02	5.92479E − 02	4.25448E − 02	2.79951E − 02	1.73610E − 02	1.02782E − 02	5.89420E − 03
	0.48	0.48	0.60	0.69	0.76	0.80	0.84
2^{-10}	8.36757E − 02	5.99848E − 02	4.29529E − 02	2.82990E − 02	1.75626E − 02	1.03986E − 02	5.96867E − 03
	0.48	0.48	0.60	0.69	0.76	0.80	0.84
2^{-14}	8.37502E − 02	6.00286E − 02	4.29776E − 02	2.83178E − 02	1.75751E − 02	1.04061E − 02	5.97332E − 03
	0.48	0.48	0.60	0.69	0.76	0.80	0.84
2^{-18}	8.37548E − 02	6.00313E − 02	4.29791E − 02	2.83190E − 02	1.75759E − 02	1.04065E − 02	5.97361E − 03
	0.48	0.48	0.60	0.69	0.76	0.80	0.84

Table 2 Maximum pointwise errors ($E_{N,\epsilon}$) and experimental order of convergence (EOC), when applied to Example 1 for various values of ϵ and N and $\delta = \varepsilon/10,\ \eta = 0$

ϵ	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$	$N = 512$	N =1,024
2^{-2}	2.72141E − 02	1.49540E − 02	7.87732E − 03	4.04774E − 03	2.05287E − 03	1.03385E − 03	5.18803E − 04
	0.86	0.92	0.96	0.98	0.99	0.99	1.0
2^{-6}	8.24516E − 02	5.92836E − 02	4.25680E − 02	2.80136E − 02	1.73735E − 02	1.02856E − 02	5.89884E − 03
	0.48	0.60	0.69	0.76	0.80	0.84	
2^{-10}	8.36772E − 02	5.99870E − 02	4.29543E − 02	2.83002E − 02	1.75633E − 02	1.03990E − 02	5.96896E − 03
	0.48	0.48	0.60	0,69	0.76	0.80	.84
2^{-14}	8.37503E − 02	6.00287E − 02	4.29777E − 02	2.83179E − 02	1.75751E − 02	1.04061E − 02	5.97334E − 03
	0.48	0.48	0.60	0.69	0.76	0.80	0.84
2^{-18}	8.37548E − 02	6.00313E − 02	4.29791E − 02	2.83190E − 02	1.75759E − 02	1.04065E − 02	5.97361E − 03
	0.48	0.48	0.60	0.69	0.76	0.80	0.84

Table 3 Maximum pointwise errors ($E_{N,\epsilon}$) and experimental order of convergence (EOC), when applied to Example 1 for various values of ϵ and N and $\delta = 0,\ \eta = 0$

ϵ	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$	$N = 512$	N =1,024
2^{-2}	2.68749E − 02	1.47519E − 02	7.75998E − 03	3.98604E − 03	2.02078E − 03	1.01751E − 03	5.10560E − 04
	0.86	0.93	0.96	0.98	0.99	0.99	1.0
2^{-6}	8.24262E − 02	5.92481E − 02	4.25449E − 02	2.79952E − 02	1.73611E − 02	1.02782E − 02	5.89422E − 03
	0.48	0.48	0.60	0.69	0.76	0.80	0.84
2^{-10}	8.36757E − 02	5.99848E − 02	4.29529E − 02	2.82990E − 02	1.75626E − 02	1.03986E − 02	5.96867E − 03
	0.48	0.48	0.60	0.69	0.76	0.80	0.84
2^{-14}	8.37502E − 02	6.00286E − 02	4.29776E − 02	2.83178E − 02	1.75751E − 02	1.04061E − 02	5.97332E − 03
	0.48	0.48	0.60	0.69	0.76	0.80	0.84
2^{-18}	8.37548E − 02	6.00313E − 02	4.29791E − 02	2.83190E − 02	1.75759E − 02	1.04065E − 02	5.97361E − 03
	0.48	0.48	0.60	0.69	0.76	0.80	0.84

convergence of the proposed scheme is proved by carrying out numerical experiments. Tables 1–4 shows maximum pointwise error and experimental rate of convergence for the considered example. It is evident from these tables that the proposed scheme converges uniformly independent of the perturbation parameter and the shift arguments.

Table 4 Maximum pointwise errors ($E_{N,\epsilon}$) and experimental order of convergence (EOC), when applied to Example 1 for various values of ϵ and N and $\delta = 0,\ \eta = \varepsilon/5$

ϵ	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$	$N = 512$	N =1,024
2^{-2}	2.71650E − 02	1.49256E − 02	7.86139E − 03	4.03942E − 03	2.04857E − 03	1.03167E − 03	5.17703E − 04
	0.86	0.92	0.96	0.98	0.99	0.99	1.0
2^{-6}	8.24514E − 02	5.92835E − 02	4.25679E − 02	2.80136E − 02	1.73734E − 02	1.02855E − 02	5.89884E − 03
	0.48	0.48	0.60	0.69	0.76	0.80	0.84
2^{-10}	8.36772E − 02	5.99870E − 02	4.29543E − 02	2.83002E − 02	1.75633E − 02	1.03990E − 02	5.96896E − 03
	0.48	0.48	0.60	0.69	0.76	0.80	0.84
2^{-14}	8.37503E − 02	6.00287E − 02	4.29777E − 02	2.83179E − 02	1.75751E − 02	1.04061E − 02	5.97334E − 03
	0.48	0.48	0.60	0.69	0.76	0.80	0.84
2^{-18}	8.37548E − 02	6.00313E − 02	4.29791E − 02	2.83190E − 02	1.75759E − 02	1.04065E − 02	5.97361E − 03
	0.48	0.48	0.60	0.69	0.76	0.80	0.84

Acknowledgements The first author thanks Council for Scientific and Industrial Research, New Delhi, India, for providing financial support.

References

1. Joseph, D. D., Preziosi, L.: Heat waves. Rev. Modern Phys. 61, 41–73 (1989).
2. Ezzat, M. A., Othman, M. I., El-Karamany, A. M. S.: State space approach to two-dimensional generalized thermo- viscoelasticity with two relaxation times. Int. J. Eng. Sci. 40, 1251–1274 (2002).
3. Doolan, E. P., Miller, J. J. H., Schilders, W. H. A.: Uniform Numerical Methods for Problems with Initial and Boundary Layers, Boole Press, Dublin (1980).
4. Elsgolts, E. L.: Qualitative methods in mathematical analysis. Translations of Mathematical Monographs, 12, American Mathematical Society, Providence, RI, 1964.
5. Liu, Q., Wang, X., De Kee, D.: Mass transport through swelling membranes. Int. J. Eng. Sci. 43, 1464–1470 (2005).
6. Longtin, A., Milton, J.: Complex oscillations in the human pupil light reflex with mixed and delayed feedback. Math. Biosci. 90, 183–199 (1988).
7. Derstine, M. W., Gibbs, H. M., Kaplan, D. L.: Bifurcation gap in a hybrid optical system. Physics Review A. 26, 3720–3722 (1982).
8. Wazewska-Czyzewska, M., Lasota, A.: Mathematical models of the red cell system. Mat. Stos. 6, 25–40 (1976).
9. Tuckwell, H. C.: On the first exit time problem for temporally homogeneous Markov Processes. J. Appl. Prob. 13, 39–48 (1976).
10. Lange, C. G., Miura, R. M.: Singular perturbation analysis of boundary value problems for differential-difference equations. V. Small shifts with layer behavior. SIAM J. Appl. Math. 54, 249–272 (1994).
11. Lange, C. G., Miura, R. M.: Singular perturbation analysis of boundary value problems for differential-difference equations. VI. Small shifts with rapid oscillations. SIAM J. Appl. Math. 54, 273–283 (1994).
12. Lange, C. G., Miura, R. M.: Singular perturbation analysis of boundary value problems for differential-difference equations. III. Turning point problems. SIAM J. Appl. Math. 45, 708–734 (1985).
13. Miller, J. J. H., O'Riordan, R. E., Shishkin, G. I.: Fitted Numerical Methods for Singular Perturbation Problems, World Scientific, Singapore, 1996.
14. Kadalbajoo, M. K., Sharma, K. K.: Numerical analysis of singularly perturbed delay differential equations with layer behavior. Appl. Math. Comp. 157, 11–28 (2004).

15. Kadalbajoo, M. K., Sharma, K. K.: ε- uniformly convergent non-standard finite difference methods for singularly perturbed differential difference equations with small delay. Appl. Math. Comp. 175, 864–890 (2006).
16. Kadalbajoo, M. K., Sharma, K. K.: A numerical method based on finite difference for boundary value problems for singularly perturbed delay differential equations. Appl. Math. Comp. 197(2), 692–707 (2008).
17. Kadalbajoo, M. K., Sharma, K. K.: Numerical treatment for singularly perturbed non-linear differential-difference equations with negative shifts. Nonlinear Anal.Th. Math. Appl. 63 (5–7), 1909–1924 (2005).
18. Kadalbajoo, M. K., Sharma, K. K.: Numerical treatment of mathematical model arising from a model of neuronal variability. J. Math. Anal. Appl. 307(2), 606–627 (2005).
19. Kadalbajoo, M. K., Sharma, K. K.: ε uniform fitted mesh method for singularly perturbed differential-difference equations: Mixed type of shifts with layer behavior. Inter. J. Comp. Math. 81(1), 49–62 (2004).
20. Mackey, M. C., Glass, L.: Oscillation and chaos in physiological control systems. Science. 197, 287–289 (1977).
21. Rai, P., Sharma, K. K.: Parameter uniform numerical method for singularly perturbed differential-difference equations with interior layer. Inter. J. Comp. Math. 88(16), 3416–3435 (2011).
22. Rai, P., Sharma, K. K.: Numerical analysis of singularly perturbed delay differential turning point problem. Appl. Math. Comp. 218 3483–3498 (2011).

Stability of Difference Schemes on Uniform Grids for a Singularly Perturbed Convection-Diffusion Equation

G. Shishkin

Abstract For a model Dirichlet problem to a singularly perturbed ordinary differential convection-diffusion equation, we discuss a "standard" approach to the construction of difference schemes that use standard grid approximations on uniform grids, the step-size of which is chosen sufficiently small for small values of a perturbation parameter ε, $\varepsilon \in (0, 1]$. It is shown that such a scheme, under its convergence in the maximum norm theoretically proved, is not ε-uniformly stable to perturbations in the data of the discrete problem. When perturbations take place and the parameter ε decreases, the actual accuracy of the computed solutions *may deteriorate up to a full accuracy loss* for sufficiently small values of ε, namely, under the condition $t = \mathcal{O}(\ln \varepsilon^{-1})$, where t is the number of computer word digits.

1 Introduction

At present, for solving singularly perturbed boundary value problems with accuracy independent of the perturbation parameter ε, grid numerical methods are sufficiently well developed that use special grids condensing in boundary layers, in particular, piecewise-uniform grids (see, e.g., [1,5,6,8,9] and the bibliography therein). In [4], a technique for experimental study of special difference schemes is developed. A "drawback" of such special numerical methods is the necessity to use nonuniform grids whose step-size sharply changes in a neighborhood of the boundary layer.

Quite often for solving problems with boundary layers, as an alternative approach, "standard" difference schemes are used, i.e., standard grid approximations of the problem on uniform grids in which the step-size is chosen sufficiently

G. Shishkin (✉)
Institute of Mathematics and Mechanics, Russian Academy of Sciences, Ekaterinburg, Russia
e-mail: shishkin@imm.uran.ru

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_32, © Springer-Verlag Berlin Heidelberg 2013

small for small values of the perturbation parameter ε, $\varepsilon \in (0, 1]$. When the step-size is much less than the parameter ε, such schemes converge but formally, i.e., only in the case when computations are performed precisely. The presence of perturbations in the computations brings to serious difficulties because the "standard" difference schemes lose stability.

The aim of this research is, with an example of an one-dimensional convection-diffusion problem, to study difficulties that arise when singularly perturbed convection-diffusion problems are solved numerically using "standard" difference schemes. Such problems were not considered before.

2 Classical Difference Scheme for a Convection-Diffusion Problem

On the set $\overline{D}$,

$$\overline{D} = D \bigcup \Gamma, \quad D = (0, d), \tag{1}$$

we consider the Dirichlet problem for the singularly perturbed ordinary differential convection-diffusion equation

$$L\,u(x) \equiv \Big\{\varepsilon a(x)\frac{d^2}{dx^2} + b(x)\frac{d}{dx} - c(x)\Big\}u(x) = f(x), \quad x \in D, \tag{2}$$

$$u(x) = \varphi(x), \quad x \in \Gamma.$$

Here $\Gamma = \Gamma_1 \cup \Gamma_2$, Γ_1 and Γ_2 are the left and right parts of the boundary Γ; the functions $a(x)$, $b(x)$, $c(x)$, $f(x)$ are assumed to be sufficiently smooth on $\overline{D}$, moreover, $a_0 \le a(x) \le a^0$, $b_0 \le b(x) \le b^0$, $c_0 \le c(x) \le c^0$, $x \in \overline{D}$; a_0, b_0, $c_0 \ge m > 0$; $|f(x)| \le M$, $x \in \overline{D}$; $|\varphi(x)| \le M$, $x \in \Gamma$; $m \le d \le M$; the parameter ε takes arbitrary values in the open-closed interval $(0, 1]$; and the constants M, m do not depend on the parameter ε.

For small values of the parameter ε, a boundary layer appears in a neighborhood of the set Γ_1.

We consider a difference scheme constructed on the basis of classical approximations of problem (2), (1) on a uniform grid.

On the set $\overline{D}$, we introduce the uniform grid

$$\overline{D}_h = \overline{D}_h^u = \overline{\omega}, \tag{3}$$

where $\overline{\omega}$ is a uniform mesh on the interval $[0, d]$ with the step-size $h = d/N$, where $N + 1$ is the number of nodes $x = x^i$ in the mesh $\overline{\omega}$, $i = 0, 1, \ldots, N$.

We approximate problem (2), (1) by the difference scheme [7]

$$\Lambda z(x) \equiv \{\varepsilon a(x)\delta_{\widehat{\overline{x}x}} + b(x)\delta_x - c(x)\}z(x) = f(x),\ x \in D_h,\ z(x) = \varphi(x),\ x \in \Gamma_h. \tag{4}$$

Here $D_h = D \bigcap \overline{D}_h$, $\Gamma_h = \Gamma \bigcap \overline{D}_h$, $\delta_{\widehat{\overline{x}x}} z(x)$ is the central difference derivative of the second order on the uniform grid, $\delta_{\widehat{\overline{x}x}} z(x) = h^{-1}[\delta_x z(x) - \delta_{\overline{x}} z(x)]$, $x = x^i \in D_h$, $i = 0, 1, \ldots, N$; $\delta_x z(x)$ and $\delta_{\overline{x}} z(x)$ are the first-order (forward and backward) difference derivatives.

The scheme (4) is ε-uniformly monotone [7]. For the solution of the scheme, using the maximum principle, we get the estimate

$$\| u - z \|_{\overline{D}_h} \leq M \left(\varepsilon + N^{-1}\right)^{-1} N^{-1}. \tag{5}$$

The scheme (4), (3) converges only under the condition

$$N^{-1} = o(\varepsilon). \tag{6}$$

Remark 1. For the convergence of scheme (4), (3) with the prescribed accuracy $\mathcal{O}(\delta)$, it is necessary to use grids with the number of nodes $N \gtrsim \varepsilon^{-1}\,\delta^{-1}$ growing unboundedly as $\varepsilon \to 0$. □

3 Conditioning of Difference Scheme (4), (3)

Here we discuss conditioning of the matrix in difference scheme (4), (3) and give estimates for perturbations of the discrete solution generated by perturbations in the data of the discrete problem.

3.1. Consider conditioning of the matrix for difference scheme (4), (3).

Write scheme (4), (3) as a system of algebraic equations. Let an $(N+1)$-dimensional vector Y correspond to $N+1$ components of the function $z(x)$ in the nodes $x \in \overline{D}_h$. Ordering the elements $z(x)$, $x \in \overline{D}_h$, in scheme (4), (3), we come to the system

$$A\,Y = F. \tag{7}$$

Here A is the tridiagonal $(N+1) \times (N+1)$-matrix (a_{ij}), F is the $(N+1)$-dimensional vector; Y and F are vectors in the normalized space R^{N+1} with the uniform vector-norm $\|\cdot\|$. The components of the matrix A and vectors Y and F in (7) are defined by the relations

$$a_{11} = 1,\ \ a_{i,i-1} = -\varepsilon\, h^{-2}\, a(x_i),\ \ a_{ii} = 2\,\varepsilon\, h^{-2}\, a(x_i) + h^{-1}\, b(x_i) + c(x_i),$$

$$a_{i,i+1} = -\varepsilon\, h^{-2}\, a(x_i) - h^{-1}\, b(x_i),\ \ i = 2, \ldots, N,\ \ a_{N+1,\,N+1} = 1;$$

$$Y_i = z(x_i),\ i = 1, \ldots, N+1;\ \ F_1 = \varphi(x_1),$$
$$F_i = -f(x_i),\ i = 2, \ldots, N,\ \ F_{N+1} = \varphi(x_{N+1});$$

here $x_{i(7)} = x^{i-1}_{(3)}$, $x^i \in \overline{D}_h$, and h is the step-size in the uniform grid $\overline{D}_h$.

The matrix A is the M-matrix with the nonstrict diagonal predominance (for $c(x) \geq 0$) and with the strict predominance in the first and last rows. The notation $Y \geq 0$ means that $Y_i \geq 0$ for all i.

The operator $A_{(7)}$ satisfies the monotonicity principle: it follows from the condition $A\,Y^1 \geq A\,Y^2$ that $Y^1 \geq Y^2$.

Using the majorant function technique (see, e.g., [5,8,9]) for the discrete problem (4), (3), we establish the ε-uniform boundedness of the norm to the matrix A^{-1} : $\|A^{-1}\| \leq M$. Here $\|A^{-1}\| \leq M$ is the matrix norm induced by the uniform vector-norm $\|\cdot\|$.

For the condition number $æ_M(A)$ of the matrix $A_{(7)}$ on the uniform grid $\overline{D}_{h(3)}$, where $æ_M(A) = \| A \| \| A^{-1} \|$, we have the unimprovable (in ε and N) estimate

$$æ_M(A;\ \overline{D}^u_{h(3)}) \leq MN\ (1 + \varepsilon\, N) \tag{8}$$

(see [9], Chap. 12 and the bibliography therein). We call ε and N the *primitive variables*.

Remark 2. *The condition number of the matrix* A on the uniform grid $\overline{D}_{h(3)}$, written in the *primitive variables* ε and N, turns out to be *ε-uniformly bounded*. But, difference scheme (4), (3) converges with the estimate (5) and only under condition (6), i.e., for $N^{-1} \ll \varepsilon$, $\varepsilon \in (0, 1]$. Thus, the estimate (8) for the condition number of the matrix A in the variables ε, N is *noninformative*, because it does not provide any relation between the matrix condition number and the convergence of the scheme (4), (3). □

It is of interest to estimate the condition number of the matrix A depending on the value δ, i.e., the accuracy of the solution to difference scheme (4), (3), and the value of the parameter ε. We call δ, ε the *informative variables*.

We estimate the condition number of the matrix A for scheme (4), (3), assuming its convergence. Let the solution of the scheme converge with the estimate (5), i.e.,

$$\| u - z \|_{\overline{D}_h} \leq M\,\delta, \tag{9}$$

where $\delta = \delta(\varepsilon,\ N) = (\varepsilon + N^{-1})^{-1}\, N^{-1}$. Then, for the matrix A and its condition number, we get the unimprovable (in ε and δ) estimate

$$\| A(\overline{D}_{h(3)}) \|,\ æ_M(A;\ \overline{D}^u_{h(3)}) \leq M\,\varepsilon^{-1}\,\delta^{-2}. \tag{10}$$

Thus, for the scheme convergent on the uniform grid with the estimate (9), the condition number of the matrix A in the variables ε, δ is not ε-uniformly bounded; the matrix A is *not ε-uniformly well conditioned*.

Remark 3. The approach based on the condition number of the matrix does not allow to clarify the ε-uniform stability of the scheme on uniform grids to perturbation in the data of the scheme that requires additional study, in particular, on the basis of *conditioning of difference scheme* (4), (3) (see, e.g., [2, 3]). □

3.2. Now we discuss the influence of perturbations in the data of difference scheme (4), (3) on perturbation of its solution depending on the parameter ε. Consider the following perturbed problem corresponding to problem (7):

$$A^* Y^* = F^*. \tag{11}$$

Here A^* is the perturbed matrix (a^*_{ij}), where Y^* and F^* are the perturbed vectors,

$$A^* = A + \delta A, \quad Y^* = Y + \delta Y, \quad F^* = F + \delta F.$$

The perturbations of the coefficient $a(x_i)$ involved in the components a_{ij}, $j = i-1, i, i+1$, $i = 2, \ldots, N$, of the matrix A, in general, are different; we denote these perturbations in the components a_{ij} by δa_i^j. In a similar way, we denote perturbations of the coefficient $b(x_i)$ in the components b_{ij}, $j = i, i+1$, by δb_i^j, and perturbations of the coefficient $c(x_i)$ in the component c_{ii} are denoted by δc_i^i; $i = 2, \ldots, N$. Assume that the components, which are equal to 0 or 1, and also the values ε and h are not perturbed. Thus, in the componentwise form of the matrix δA and vectors δF and δY, we have

$$\delta a_{11} = \delta a_1^1 = 0,\ \delta a_{i,i-1} = -\varepsilon h^{-2} \delta a_i^{i-1},\ \delta a_{ii} = 2\varepsilon h^{-2} \delta a_i^i + h^{-1} \delta b_i^i + \delta c_i^i,$$

$$\delta a_{i,i+1} = -\varepsilon h^{-2} \delta a_i^{i+1} - h^{-1} \delta b_i^{i+1},\ i = 2, \ldots, N,\ \delta a_{N+1,N+1} = \delta a_{N+1}^{N+1} = 0; \tag{12}$$

$$\delta F_1 = \delta \varphi(x_1), \quad \delta F_i = -\delta f(x_i), \quad i = 2, \ldots, N,$$

$$\delta F_{N+1} = \delta \varphi(x_{N+1}); \quad \delta Y_i = \delta z(x_i), \quad i = 1, \ldots, N+1.$$

With regard to the ε-uniform boundedness of $\| Y \|$, we find the estimate

$$\| \delta Y \| \leq M \left(\| \delta F \| + \| \delta A \| \right). \tag{13}$$

For the perturbation of the discrete solution $z^*(x) - z(x)$, taking into account relations (12) and (13), we obtain the estimate

$$\| z^* - z \|_{\overline{D}_h} \leq M \left[\varepsilon N^2 \max_{i,j;\, i \neq 1, N+1} |\delta a_i^j| + N \max_{i,j} |\delta b_i^j| + \max_i \widehat{\psi}_i^i \right], \tag{14}$$

where $\max\limits_i \widehat{\psi}_i^i = \max \left[\max\limits_{i;\, i=1, N+1} |\delta a_i^i|,\ \max\limits_i |\delta c_i^i|,\ \max\limits_i |\delta f_i|,\ \max\limits_i |\delta \varphi_i| \right]$,

δf_i are the perturbations $\delta f(x_i)$ for $i = 2, \ldots, N$ and $\delta \varphi_i$ are the perturbations $\delta \varphi(x_i)$ for $i = 1,\ N+1$; the estimate is unimprovable in ε, N.

Taking into account estimates (5) and (14), for the function $z^*(x)$, i.e., the solution of a perturbed difference scheme corresponding to matrix notation (11),

we get the estimate (unimprovable in ε, N)

$$\| u - z^* \|_{\overline{D}_h} \le M[(\varepsilon + N^{-1})^{-1} N^{-1} + \tag{15}$$

$$\varepsilon N^2 \max_{i,j;i\neq 1,N+1} |\delta a_i^j| + N \max_{i,j} |\delta b_i^j| + \max_i \widehat{\psi}_i^i].$$

For perturbed problem (11), we have the following perturbed difference scheme

$$\Lambda^* z^*(x) \equiv \left\{\varepsilon\, a^*(x)\, \delta_{\widehat{xx}} + b^*(x)\, \delta_x - c^*(x)\right\} z^*(x) = f^*(x),\ x \in D_h, \qquad z^*(x) = \varphi^*(x),\ x \in \Gamma_h. \tag{16a}$$

Here $x = x^i$, $x^i \in \overline{D}_{h(3)}$, $i = 0, 1, 2, \ldots, N$, and in (16b) $x_i = x_i(x^{i-1}) = x_{(3)}^{i-1}$, $x^i \in \overline{D}_{h(3)}$

$$a^*(x_i) = a(x_i) + \delta a_i^{i-1},\ \ b^*(x_i) = b(x_i) + \delta b_i^{i+1} + \varepsilon\, h^{-1}\,(-\delta a_i^{i-1} + \delta a_i^{i+1}), \tag{16b}$$

$$c^*(x_i) = c(x_i) + \delta c_i^i - \varepsilon h^{-2}(\delta a_i^{i-1} - 2\delta a_i^i + \delta a_i^{i+1}) + h^{-1}(\delta b_i^i - \delta b_i^{i+1}),\ \ x^{i-1} \in D_h;$$

$$f^*(x_i) = f(x_i) + \delta f_i\ \ x^{i-1} \in D_h;\ \ \varphi^*(x_i) = \varphi(x_i) + \delta\varphi_i\ \ x^{i-1} \in \Gamma_h.$$

It follows from estimate (15) that, when the perturbations in the data sufficiently rapidly tend to zero as N grows, the solution of perturbed difference scheme (16), (3) converges to the solution of nonperturbed difference scheme (4), (3) and, hence, to the solution of boundary value problem (2), (1) under the unimprovable condition

$$\varepsilon^{-1} N^{-1} = o(1), \quad \max_{i,j;i\neq 1,N+1} |\delta a_i^j| = o(\varepsilon^{-1} N^{-2}), \quad \max_{i,j} |\delta b_i^j| = o(N^{-1}), \tag{17}$$

$$\max_{i;i=1,N+1} |\delta a_i^i|, \quad \max_i |\delta c_i^i|, \quad \max_i |\delta f_i|, \quad \max_i |\delta\varphi_i| = o(1); \quad N \to \infty, \quad \varepsilon \in (0, 1].$$

3.3. In the case of difference scheme (4), (3), we write out an estimate for the perturbation of the discrete solution $z^*(x) - z(x)$ and an estimate for the error of the perturbed discrete solution $u(x) - z^*(x)$, using the informative variables ε, δ.

With regard to (14), we have

$$\| z^* - z \|_{\overline{D}_h} \le M[\varepsilon^{-1}\delta^{-2} \max_{i,j;i\neq 1,N+1} |\delta a_i^j| + \varepsilon^{-1}\delta^{-1} \max_{i,j} |\delta b_i^j| + \max_i \widehat{\psi}_i^i], \tag{18}$$

where $\max_i \widehat{\psi}_i^i = \max_i \widehat{\psi}_{i\,(14)}^i$, $\delta = \delta_{(9)}(\varepsilon, N)$; the estimate is unimprovable in ε, δ.

Definition 1. It is convenient to write the unimprovable estimate (18) in *informative variables* ε, δ as the estimate of the *relative error* $\|\delta z_u\|_{\overline{D}_h} / \|z_u\|_{\overline{D}_h}$, using *relative perturbations* in the data of the grid solution in the matrix form (11),

$$\frac{\|\delta\, z_u\,\|_{\overline{D}_h}}{\|z_u\,\|_{\overline{D}_h}} \;\le\; æ_P(A;\overline{D}_h)\,\Big(\frac{\|\delta\, F\,\|}{\|F\,\|} \;+\; \frac{\|\delta\, A\|}{\|A\|}\Big).$$

We call the quantity $æ_P(A;\overline{D}_h)$ *the condition number of the grid problem*, i.e., of difference scheme (4) on uniform grid (3) (see discussions of conditioning for a matrix and for a problem in [2, 3]). □

With regard to estimate (18), we obtain the estimate of $æ_P(A;\overline{D}_h)$ in informative variables

$$æ_P(A;\overline{D}_h) \le M\,\varepsilon^{-1}\,\delta^{-2}. \tag{19}$$

Theorem 1. *Difference scheme* (4), (3) *is not* ε*-uniformly well conditioned. For the condition number* $æ_P(A;\overline{D}_h)$ *of the difference scheme convergent at the rate* $\mathcal{O}(\delta)$*, the estimate* (19) *holds.*

Remark 4. Bound (19) with respect to orders of the values δ and ϵ is achievable, i.e.,

$$|æ_P(A;\overline{D}_h)| \lesssim M\,\varepsilon^{-1}\,\delta^{-2}.$$

Under the condition $a(x) = 1$, we obtain the bound $æ_P(A;\overline{D}_h) \lesssim M\,\varepsilon^{-1}\,\delta^{-1}$; for $a(x) = b(x) = 1$, we have $æ_P(A;\overline{D}_h) \lesssim M$. In these bounds, $\delta = \delta_{(19)}(\varepsilon, N)$. □

4 Stability of Difference Scheme (4), (3) to Perturbation in the Data

In the case of difference scheme (4), (3), we discuss the behaviour of both the perturbation $z^*(x) - z(x)$ of the grid solution depending on the informative variables ε, δ, and the perturbation in the data of the scheme.

Definition 2. Assume that the solution of difference scheme (4), (3) converges, as $N \to \infty$, with the estimate

$$\|\, u - z\,\|_{\overline{D}_h} \le M\,\delta, \quad \varepsilon \in (0, 1].$$

We say that *scheme* (4), (3) is *stable to perturbation* in the data *for fixed values of the parameters* ε *and* δ, if the perturbation $\|z^* - z\|_{\overline{D}_h}$ satisfies the estimate

$$\|z^* - z\|_{\overline{D}_h} \le Q(\varepsilon,\,\delta)\,\lambda\big(\max_{i,j}|\delta a_i^j|,\,\max_{i,j}|\delta b_i^j|,\,\max_i \widehat{\psi}_i^i\,\big),$$

where $\lambda(\xi_1, \xi_2, \xi_3) \to 0$ for $\xi_1, \xi_2, \xi_3 \to 0$; $\max_i \widehat{\psi}_i^i = \max_i \widehat{\psi}_{i\,(14)}^i$; and $Q(\varepsilon, \delta)$ is the coefficient depending on ε, δ. Let the following estimate hold for $\|z^* - z\|_{\overline{D}_h}$:

$$\|z^* - z\|_{\overline{D}_h} \le M\, \lambda(\, \varepsilon^{-\alpha_1}\, \delta^{-\beta_1} \max_{i,j;i\ne 1,N+1} |\delta a_i^j|,\ \varepsilon^{-\alpha_2}\, \delta^{-\beta_2} \max_{i,j} |\delta b_i^j|,\ \max_i \widehat{\psi}_i^i\,), \tag{20}$$

where $\alpha_1, \alpha_2, \beta_1, \beta_2 \ge 0$. We say that *scheme* (4), (3) is *ε-uniformly stable to perturbation* in the data if $\alpha_1, \alpha_2 = 0$ in estimate (20). □

With regard to estimate (18), the *scheme* (4), (3) is *not ε-uniformly stable to perturbation in the data*, in particular, to perturbation in computations.

The unimprovable estimate (15) implies the estimate

$$\|u - z^*\|_{\overline{D}_h} \le M\, [\, \delta + \varepsilon^{-1}\, \delta^{-2} \max_{i,j;i\ne 1,N+1} |\delta a_i^j| + \varepsilon^{-1}\, \delta^{-1} \max_{i,j} |\delta b_i^j| + \max_i \widehat{\psi}_i^i\,], \tag{21}$$

unimprovable in ε and δ; $\delta = \delta_{(9)}(\varepsilon, N)$.

Theorem 2. *Difference scheme* (4), (3) *is not ε-uniformly stable to perturbation in the data. The solution of perturbed scheme* (16), (3) *satisfies estimate* (21).

With regard to estimate (21), the following theorem is established:

Theorem 3. *The condition imposed on the perturbations*

$$\max_{i,j;i\ne 1,N+1} |\delta a_i^j| = o(\varepsilon\, \delta^2), \quad \max_{i,j} |\delta b_i^j| = o(\varepsilon\, \delta), \tag{22}$$

$$\max_{i;i=1,N+1} |\delta a_i^i|,\ \max_i |\delta c_i^i|,\ \max_i |\delta f_i|,\ \max_i |\delta \varphi_i| = o(1);\ \ \delta \to 0,\ \ \varepsilon \in (0, 1],$$

is necessary and sufficient for the convergence of the perturbed difference scheme (16), (3).

Remark 5. The absence of the ε-uniform stability to perturbation in the data of standard scheme (4) on *uniform grid* (3) rather *restricts the usefulness* of such schemes because, when solving the problems on a computer with a fixed number t in the computer word digits, the *accuracy* of the computed solutions *comes down* as the parameter ε decreases; moreover, the accuracy *deteriorates up to a full accuracy loss* for small values of ε, namely, under the condition $t = \mathcal{O}(\ln \varepsilon^{-1})$.

With the use of scheme (4), (3) for solving of *singularly perturbed boundary value problem* (2), (1) with the prescribed accuracy δ for $\varepsilon \to 0$, besides the well known problem noted in Remark 1 (the *necessity to increase unboundedly N, i.e., the number of mesh nodes*, as $N \gtrsim \varepsilon^{-1}\, \delta^{-1}$), a new problem appears, i.e., the *necessity* to use a computer with the *number of the computer word digits that grows unboundedly*, namely, with $t \gtrsim \ln(\varepsilon^{-1} + \delta^{-1})$. □

Acknowledgements This research was supported by the Russian Foundation for Basic Research under grant No. 10-01-00726.

References

1. Bakhvalov, N.S.: On the optimization of methods for solving boundary value problems in the presence of a boundary layer. USSR Comput. Maths. Math. Phys. **9**. 139–166 (1969)
2. Bakhvalov, N.S.: Numerical Methods. Nauka, Moscow (in Russian) (1973)
3. Bakhvalov, N.S., Zhidkov, N.P., Kobelkov, G.M.: Numerical Methods. LBZ, Moscow (in Russian) (2001)
4. Farrell, P.A., Hegarty, A.F., Miller, J.J.H., O'Riordan, E., Shishkin, G.I.: Robust Computational Techniques for Boundary Layers. Chapman and Hall/CRC, Boca Raton (2000)
5. Miller, J.J.H., O'Riordan, E., Shishkin G.I.: Fitted numerical methods for singular perturbation problems. World Scientific, Singapore (1996)
6. Roos, H.-G., Stynes, M., Tobiska L.: Numerical Methods for Singularly Perturbed Differential Equations. Convection-Diffusion-Reaction and Flow Problems. In: Springer Series in Computational Mathematics **24**. Springer-Verlag, Berlin, 2-nd edn. (2008)
7. Samarskii, A.A.: The Theory of Difference Schemes. Marcel Dekker, Inc., New York (2001)
8. Shishkin, G.I.: Discrete Approximations of Singularly Perturbed Elliptic and Parabolic Equations. Ural Branch of Russian Academy of Sciences, Ekaterinburg (1992)
9. Shishkin, G.I., Shishkina, L.P.: Difference Methods for Singular Perturbation Problems. In: Chapman & Hall/CRC Monographs and Surveys in Pure and Applied Mathematics. **140**. CRC Press, Boca Raton (2009)

Difference Scheme of the Solution Decomposition Method for a Singularly Perturbed Parabolic Convection-Diffusion Equation

L. Shishkina and G. Shishkin

Abstract For a Dirichlet problem for an one-dimensional singularly perturbed parabolic convection-diffusion equation, a difference scheme of the solution decomposition method is constructed. This method involves a special decomposition based on the asymptotic construction technique in which the regular and singular components of the grid solution are solutions of grid subproblems solved on *uniform grids*, moreover, the coefficients of the grid equations do not depend on the singular component of the solution unlike the fitted operator method. The constructed scheme converges in the maximum norm ε-uniformly (i.e., independent of a perturbation parameter ε, $\varepsilon \in (0, 1]$) at the rate $\mathcal{O}\left(N^{-1} \ln N + N_0^{-1}\right)$ the same as a scheme of the condensing grid method on a piecewise-uniform grid (here N and N_0 define the numbers of the nodes in the spatial and time meshes, respectively).

1 Introduction

At present, for singularly perturbed boundary value problems, methods for constructing difference schemes convergent in the maximum norm independently of a perturbation parameter ε, $\varepsilon \in (0, 1]$ (we say, ε-uniformly) are well developed. They are methods used piecewise-uniform grids condensing in a neighborhood of the boundary layer (see, e.g., [1, 10, 11]) and fitted operator methods (their description see, e.g., in [2–4, 7] and the bibliography therein. The condensing grids methods have widespread application due to their simplicity. A drawback of these numerical methods is the necessity to solve discrete equations on grids in which step-sizes change sharply in a neighborhood of the boundary layer that gives rise to difficulties in the construction of high-order accurate schemes and also in approximation of derivatives. Fitted operator methods have advantage in the simplicity of uniform

L. Shishkina (✉) · G. Shishkin
Institute of Mathematics and Mechanics, Russian Academy of Sciences, Ekaterinburg, Russia
e-mail: lida@convex.ru; shishkin@imm.uran.ru

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_33,

grids used, however, they have a restricted applicability, in particular, for problems with parabolic initial or boundary layers (see, e.g., [9], and also [10, 11] and the bibliography therein).

In the present paper, for a singularly perturbed parabolic convection-diffusion equation, a new approach based on the asymptotic construction technique is proposed to construct special schemes of the *solution decomposition method*, using classical approximations of subproblems on *uniform grids* for the regular and singular components of the grid solution (description of this method for an ordinary differential reaction-diffusion equation see, e.g., in [12]). The constructed difference scheme of the *solution decomposition method* converges in the maximum norm ε-uniformly with the first accuracy order up to a logarithmic factor. The new method gives opportunity to apply, for singularly perturbed problems, a technique developed in [6, 8] for regular boundary value problems. So, special *ε-uniformly convergent difference schemes of high-order accuracy* have been constructed in [13], and also *improved approximation* of solutions and *derivatives* have been achieved in [14]. Note that the standard schemes in [6, 8], when applied to singularly perturbed problems, converge only *under condition* when the step-size across the boundary layer is much less than the perturbation parameter ε.

2 Problem Formulation. Aim of Research

On the set $\overline{G}$, where $\overline{G} = G \bigcup S$, $G = D \times (0, T]$, $D = (0, d)$, we consider the Dirichlet problem for the parabolic convection-diffusion equation

$$L\,u(x,t) = f(x,t), \quad (x,t) \in G, \quad u(x,t) = \varphi(x,t), \quad (x,t) \in S. \tag{1}$$

Here[1] $L = L_{(1)} = \varepsilon\, a(x,t)\dfrac{\partial^2}{\partial x^2} + b(x,t)\dfrac{\partial}{\partial x} - c(x,t) - p(x,t)\dfrac{\partial}{\partial t}$, $(x,t) \in G$, the functions $a(x,t)$, $b(x,t)$, $c(x,t)$, $p(x,t)$, $f(x,t)$ and $\varphi(x,t)$ are assumed to be sufficiently smooth on $\overline{G}$, and S respectively, moreover,[2]

$$a_0 \le a(x,t) \le a^0, \quad b_0 \le b(x,t) \le b^0, \quad |c(x,t)| \le c^0, \quad p_0 \le p(x,t) \le p^0, \tag{2}$$

$$|f(x,t)| \le M, \quad (x,t) \in \overline{G}; \quad |\varphi(x,t)| \le M, \quad (x,t) \in S; \quad a_0,\, b_0,\, p_0 > 0;$$

the parameter ε takes arbitrary values in the open-closed interval $(0, 1]$.

[1] Notation $L_{(j.k)}$ $(\overline{G}_{(j.k)}, M_{(j.k)})$ means that this operator (domain, constant) is introduced in the formula $(j.k)$.

[2] We denote by M (by m) sufficiently large (small) positive constants that do not depend on the value of the parameter ε. In the case of grid problems, these constants are also independent of the stencils of the difference schemes.

We assume that the data of problem (1) on the set of corner points $S^* = S_0 \bigcap \overline{S}^L$ satisfy compatibility conditions guaranteeing the required smoothness of the solution on $\overline{G}$ (see, e.g., [5]). Here $S = S_0 \bigcup S^L$, S_0 and S^L are the lower and lateral parts of the boundary; $S_0 = \overline{S}_0$.

For small values of ε, a regular boundary layer appears in a neighbourhood of the set $S_1^L = \{(x,t) : \ x = 0, \ 0 < t \le T\}$. Here S_1^L and S_2^L are the left and right parts of the lateral boundary; $S^L = S_1^L \bigcup S_2^L$.

Our **aim** is for initial-boundary value problem (1), on the basis of the special decomposition of the solution into regular and singular components, to construct an ε-uniformly convergent (in the maximum norm) difference scheme using grid approximations of the components on the related uniform grids.

3 Difference Scheme on Uniform and Piecewise-Uniform Grids for Problems (1)

We consider difference schemes constructed on the basis of classical approximations of problem (1) on uniform and piecewise-uniform grids.

On the set $\overline{G}$ we introduce the rectangular grid

$$\overline{G}_h = \overline{\omega} \times \overline{\omega}_0, \tag{3}$$

where $\overline{\omega}$ and $\overline{\omega}_0$ are arbitrary, in general, nonuniform meshes on the intervals $[0,d]$ and $[0,T]$, respectively. Let $h^i = x^{i+1} - x^i$, x^i, $x^{i+1} \in \overline{\omega}$, $h = \max_i h^i$, and $h_t^k = t^{k+1} - t^k$, t^k, $t^{k+1} \in \overline{\omega}_0$, $h_t = \max_k h_t^k$. We assume the following conditions hold: $h \le M\, N^{-1}$, $h_t \le M\, N_0^{-1}$, where $N+1$ and N_0+1 are the numbers of the nodes of the meshes $\overline{\omega}$ and $\overline{\omega}_0$, respectively.

We approximate problem (1) by the monotone difference scheme [8]

$$\Lambda z(x,t) = f(x,t,), \ \ (x,t) \in G_h, \quad z(x,t) = \varphi(x,t), \ \ (x,t) \in S_h. \tag{4}$$

Here $G_h = G \bigcap \overline{G}_h$, $S_h = S \bigcap \overline{G}_h$, $\Lambda \equiv \varepsilon\, a(x,t)\, \delta_{\widehat{xx}} + b(x,t)\, \delta_x - c(x,t) - p(x,t)\, \delta_{\overline{t}}$, $(x,t) \in G_h$, $\delta_{\widehat{xx}} z(x,t) = 2(h^i + h^{i-1})^{-1}[\delta_x z(x,t) - \delta_{\overline{x}} z(x,t)]$ is the central difference derivative of the second order on a nonuniform grid, $(x,t) = (x^i,t) \in G_h$; $\delta_x z(x,t)$ and $\delta_{\overline{x}} z(x,t)$ are the first-order (forward and backward) difference derivatives.

In the case of the grid $\overline{G}_h$, uniform in both variables:

$$\overline{G}_h = \overline{G}_h^u \equiv \overline{\omega} \times \overline{\omega}_0, \tag{5}$$

using the maximum principle, we get the estimate

$$|u(x,t) - z(x,t)| \le M \left[\left(\varepsilon + N^{-1} \right)^{-1} N^{-1} + N_0^{-1} \right], \quad (x,t) \in \overline{G}_h. \tag{6}$$

The scheme (4) and (5) converges under the condition

$$N^{-1} = o(\varepsilon), \quad N_0^{-1} = o(1). \tag{7}$$

Now we construct a difference scheme on a piecewise-uniform grid that converges ε-uniformly (see, e.g., [7, 10]). On the set $\overline{G}$, we introduce the grid

$$\overline{G}_h = \overline{G}_h^* \equiv \overline{\omega}^* \times \overline{\omega}_0, \tag{8a}$$

where $\overline{\omega}_0 = \overline{\omega}_{0(5)}$, $\overline{\omega}^*$ is a piecewise-uniform mesh, which is constructed as follows. The interval $[0, d]$ is divided into the two intervals $[0, \sigma]$, $[\sigma, d]$, the step-sizes on these intervals are constant and equal to $h^{(1)} = 2\,\sigma\,N^{-1}$ and $h^{(2)} = 2(d-\sigma)N^{-1}$, respectively. The parameter σ is defined by the relation

$$\sigma = \sigma(\varepsilon,\, N,\, l) = \min\left[2^{-1}\, d,\, l\, m^{-1}\, \varepsilon\, \ln N\right], \tag{8b}$$

where $m \in (0, m_0)$, $m_0 = \min_{\overline{G}}[a^{-1}(x,t)\, b(x,t)]$; here $l = 1$. The mesh $\overline{\omega}^*$ and the mesh $\overline{G}_h^*$ are constructed.

For the solution of difference scheme (4) and (8), we obtain the estimate

$$|u(x,t) - z(x,t)| \le M \left\{ N^{-1}\, \min\left[\varepsilon^{-1},\, \ln N\right] + N_0^{-1} \right\}, \quad (x,t) \in \overline{G}_h, \tag{9}$$

and also ε-uniform estimate

$$|u(x,t) - z(x,t)| \le M \left[N^{-1}\, \ln N + N_0^{-1} \right], \quad (x,t) \in \overline{G}_h. \tag{10}$$

The scheme (4) and (8) converges ε-uniformly with the first order in t, and with the first order, up to a logarithmic factor, in x; for fixed values of ε, this scheme converges with the first order.

4 Scheme of the Solution Decomposition Method for Problem (1)

Using the decomposition of the solution to differential problem (1), we construct a difference scheme based on the discrete solution decomposition method, in which the discrete regular and singular components of the solution are computed on uniform grids.

4.1. Construct the decomposition of the solution to problem (1), and on its basis we construct a scheme of the asymptotic construction method.

We write the solution of the initial-boundary value problem as the sum of its regular and singular components

$$u(x,t) = U(x,t) + V(x,t), \quad (x,t) \in \overline{G}. \tag{11a}$$

The regular component $U(x,t)$ is represented as the expansion

$$U(x,t) = U_0(x,t) + v_U(x,t), \quad (x,t) \in \overline{G}, \tag{11b}$$

where $U_0(x,t)$ is its main term, and $v_U(x,t)$ is its remainder term.

The functions $U_0(x,t)$ and $v_U(x,t)$ in (11b) are solutions of the problems

$$L_{(12)}U_0(x,t) = f(x,t),\ (x,t) \in G \bigcup S_1^L, \quad U_0(x,t) = \varphi(x,t),\ (x,t) \in S \setminus S_1^L; \tag{12a}$$

$$L_{(1)}v_U(x,t) = -\varepsilon a(x,t)\frac{\partial^2}{\partial x^2}U_0(x,t),\ (x,t) \in G, \quad v_U(x,t) = 0,\ (x,t) \in S. \tag{12b}$$

Here $L_{(12)} \equiv b(x,t)\dfrac{\partial}{\partial x} - c(x,t) - p(x,t)\dfrac{\partial}{\partial t}$, which is the operator $L_{(1)}$ for $\varepsilon = 0$.

The function $V(x,t)$, $(x,t) \in \overline{G}$, is the solution of the problem

$$L_{(1)}V(x,t) = 0,\ (x,t) \in G, \quad V(x,t) = \varphi(x,t) - U(x,t),\ (x,t) \in S, \tag{13}$$

where $U(x,t) = U_{(11)}(x,t),\ (x,t) \in \overline{G}$.

Assume that the data of problem (1), besides the compatibility conditions on the set S^* guaranteeing the smoothness of the problem solution $u(x,t)$, satisfy additional conditions on the set S_2^* ($S_s^* = S_0 \bigcap \overline{S}_s^L$, $s = 1,2$), providing sufficient smoothness for the components of the functions $u(x,t)$. Write out these conditions

$$\begin{aligned}
&\frac{\partial^k}{\partial x^k}\varphi(x,t),\ \frac{\partial^{k_0}}{\partial t^{k_0}}\varphi(x,t) = 0,\ k + 2k_0 \le l_0,\\
&\frac{\partial^{k+k_0}}{\partial x^k\,\partial t^{k_0}}f(x,t) = 0,\ k + 2k_0 \le l_0 - 2, \quad (x,t) \in S_1^*;\\
&\frac{\partial^k}{\partial x^k}\varphi(x,t),\ \frac{\partial^{k_0}}{\partial t^{k_0}}\varphi(x,t) = 0,\ k + k_0 \le l,\\
&\frac{\partial^{k+k_0}}{\partial x^k\,\partial t^{k_0}}f(x,t) = 0,\ k + k_0 \le l - 1, \quad (x,t) \in S_2^*,
\end{aligned} \tag{14}$$

where $l = l_0 + 2$ with $l_0 > 0$ is even.

Let the data of the initial-boundary value problem (1) satisfy the condition $a, b, c, p \in C^{l_1,l_1}(\overline{G})$, $f \in C^{l_1,l_1}(\overline{G})$, $\varphi \in C^{l_1,l_1}(\overline{G})$, $l_1 = l + \alpha$, $l = l_0 + 2$, $\alpha > 0$, and also condition (14) for $l_0 = 4$. In this case for the components in the representation (11a), we receive the following estimates (see, e.g., [11]):

$$\left| \frac{\partial^{k+k_0}}{\partial x^k \partial t^{k_0}} U(x,t) \right| \le M \left[1 + \varepsilon^{2-k-k_0} + \varepsilon^{1-k-k_0} \exp(-m\,\varepsilon^{-1}\,x)\right], \tag{15a}$$

$$\left| \frac{\partial^{k+k_0}}{\partial x^k \partial t^{k_0}} V(x,t) \right| \le M \left[\varepsilon^{-k} + \varepsilon^{1-k-k_0}\right] \exp(-m\,\varepsilon^{-1}\,x), \quad (x,t) \in \overline{G}, \tag{15b}$$

where $k + 2k_0 \le K, m \in (0, m_0), m_0 = m_{(8)}$; $K = 4$.

For the components from (11b) we have the estimates

$$\left| \frac{\partial^{k+k_0}}{\partial x^k \partial t^{k_0}} U_0(x,t) \right| \le M, \quad k + k_0 \le K + 2, \tag{16a}$$

$$\left| \frac{\partial^{k+k_0}}{\partial x^k \partial t^{k_0}} v_U(x,t) \right| \le M \left[\varepsilon + \varepsilon^{2-k-k_0} + \varepsilon^{1-k-k_0} \exp(-m\,\varepsilon^{-1}\,x)\right], \tag{16b}$$

$$k + 2k_0 \le K, \qquad (x,t) \in \overline{G}.$$

4.2. Construct a difference scheme for initial-boundary value problem (1).

For not too small values of the parameter ε, namely, under the condition

$$\varepsilon \ge \varepsilon_0(N), \ \varepsilon_0(N) = m\,l^{-1}\,d\ \ln^{-1} N, \tag{17}$$

where $m = m_{(8)}$, $l = 1$, $N = N_{(5)}$, we approximate problem (1) by the standard difference scheme (4) on the uniform grid (5). Using the values of the solution $z_u(x,t)$ of the difference scheme (4) and (5), we construct the bilinear interpolant

$$\overline{z}_u(x,t), \quad (x,t) \in \overline{G} \text{ under condition (17)} \tag{18a}$$

which we call the solution of scheme {(4), (5); (17)}.

For sufficiently small values of the parameter ε, namely, under the condition

$$\varepsilon < \varepsilon_0(N) \tag{19}$$

we approximate the components in the representation (11b) and the function $U(x,t)$ on the uniform grid (5), and the singular component $V(x,t)$ from (11a) is approximated on a uniform grid constructing on the subdomain $\overline{G}^{\sigma}$ from $\overline{G}$ and adjoining to the boundary S_1^L, where

$$\overline{G}^{\sigma} = G^{\sigma} \bigcup S^{\sigma}, \quad G^{\sigma} = D^{\sigma} \times (0, T], \quad D^{\sigma} = (0, \sigma), \tag{20a}$$

$$\sigma = \sigma(\varepsilon, N, l) = \min\left[d, m^{-1}\,l\,\varepsilon\,\ln N\right]. \tag{20b}$$

Construct a difference scheme under condition (19). We approximate differential problems (12a) and (12b) by the following problems on the grid (5):

$$\Lambda_{(21)} z_{U_0}(x,t) \equiv \{b(x,t)\delta_x - c(x,t) - p(x,t)\delta_{\overline{t}}\} z_{U_0}(x,t) = f(x,t), \quad (x,t) \in G_h \bigcup S_{1h}^L,$$

$$z_{U_0}(x,t) = \varphi(x,t), \quad (x,t) \in S_h \setminus S_{1h}^L; \tag{21}$$

$$\Lambda_{(4)} z_{vU}(x,t) = -\varepsilon\, a(x,t)\, \delta_{x\overline{x}} z_{U_0}(x,t), \quad (x,t) \in G_h,$$

$$z_{vU}(x,t) = 0, \quad (x,t) \in S_h. \tag{22}$$

Here $G_h = G \bigcap \overline{G}_h$, $S_h = S \bigcap \overline{G}_h$, $S_h = S_{0h} \bigcup S_{1h}^L \bigcup S_{2h}^L$. Set

$$z_U(x,t) = z_{U_0}(x,t) + z_{vU}(x,t), \quad (x,t) \in \overline{G}_h. \tag{23}$$

By $\overline{z}_U(x,t)$, $(x,t) \in \overline{G}$ we denote the linear interpolant constructed using the values of $z_U(x,t)$ at the nodes of the grid $\overline{G}_h$. The function $z_U(x,t)$, $(x,t) \in \overline{G}_h$, and its interpolant $\overline{z}_U(x,t)$, $(x,t) \in \overline{G}$, are called, respectively, the solutions (discrete and continual) of difference scheme {(21)–(22), (5); (19)}.

Approximate problem (13). On the set $\overline{G}_{(20)}^{\sigma}$ we introduce the uniform grid

$$\overline{G}_h^{\sigma} = \overline{G}_h^{\sigma u} \equiv \overline{\omega}^{\sigma} \times \overline{\omega}_0, \tag{24}$$

where $\overline{\omega}_0 = \overline{\omega}_{0(5)}$, $\overline{\omega}^{\sigma}$ is uniform mesh on $\overline{D}_{(20)}^{\sigma}$ with the step-size $h^{\sigma} = \sigma N^{-1}$, $N+1$ is the number of nodes in the mesh $\overline{\omega}^{\sigma}$. On the grid $\overline{G}_h^{\sigma}$ we solve the discrete problem

$$\Lambda_{(4)} z_V(x,t) = 0, \quad (x,t) \in G_h^{\sigma}, \tag{25}$$

$$z_V(x,t) = \left\{ \begin{array}{cc} \varphi(x,t) - z_U(x,t), & (x,t) \in S_{1h}^{\sigma L} \\ 0, & (x,t) \in S_h^{\sigma} \setminus S_{1h}^{\sigma L} \end{array} \right\}, \quad (x,t) \in S_h^{\sigma}.$$

Using the function $z_V(x,t)$, $(x,t) \in \overline{G}_h^{\sigma}$, we construct the interpolant $\overline{z}_V(x,t)$, $(x,t) \in \overline{G}^{\sigma}$. The function $z_V(x,t)$ and $\overline{z}_V(x,t)$ outside the set $\overline{G}^{\sigma}$ are assumed equal to zero. The functions $z_V(x,t)$, $(x,t) \in \overline{G}_h^{\sigma}$, and $\overline{z}_V(x,t)$, $(x,t) \in \overline{G}^{\sigma}$, are approximations of the solution of problem (13) under condition (19). Set

$$\overline{z}_u(x,t) = \overline{z}_U(x,t) + \overline{z}_V(x,t), \quad (x,t) \in \overline{G} \ \textit{ under condition } (19). \tag{18b}$$

The function $\overline{z}_u(x,t)$ is called the solution of difference scheme {{(21)–(22), (5); (25), (24)}; (19)}, approximating problem (1) under condition (19).

The constructed function $\overline{z}_{u\,(18a,b)}(x,t)$, $(x,t) \in \overline{G}$ approximates the solution of the problem (1). This function and also the grid functions $z_{U_0}(x)$, $z_{v_U}(x,t)$, $(x,t) \in \overline{G}_h$ and $z_V(x,t)$, $(x,t) \in \overline{G}_h^{\sigma}$, are called the solutions (respectively, continual and discrete) of the difference *scheme of the solution decomposition method* {{(4), (5)}; {(21)–(22), (5); (25), (24)}}. The scheme of the solution decomposition method is ε-uniformly monotone.

Estimate $u(x,t) - \overline{z}_u(x,t)$, $(x,t) \in \overline{G}$, assuming that the estimates (15) and (16) hold. Under this, the inclusions $U, V \in C^{4,2}(\overline{G})$, and $U_0 \in C^{4,4}(\overline{G})$, $v_U \in C^{4,2}(\overline{G})$ take place. Taking into account estimates of the components $U_0(x,t)$, $v_U(x,t)$ and $V(x,t)$, we obtain estimates for $U_0(x,t) - \overline{z}_{U_0}(x,t)$, $v_U(x,t) - \overline{z}_{v_U}(x,t)$ and $V(x,t) - \overline{z}_V(x,t)$. From these estimates we get the estimate for $u(x,t) - \overline{z}_{u\,(18\,b)}(x,t)$ under condition (19). Estimating $u(x,t) - \overline{z}_{u\,(18a)}(x,t)$ under condition (17), we use estimate (6) with $\varepsilon > \varepsilon_0(N)$.

For the solution of difference scheme {{(4), (5)}; {(21)–(22), (5); (25), (24)}} of the solution decomposition method we get the following estimate:

$$|u(x,t) - \overline{z}_u(x,t)| \le M\,\{N^{-1}\min[\varepsilon^{-1},\, \ln N] + N_0^{-1}\}, \quad (x,t) \in \overline{G}, \tag{26}$$

and also the ε-uniform estimate

$$|u(x,t) - \overline{z}_u(x,t)| \le M\,[N^{-1}\,\ln N + N_0^{-1}], \quad (x,t) \in \overline{G}. \tag{27}$$

Thus, the solution to the scheme of the solution decomposition method converges ε-uniformly at the rate $\mathscr{O}\left(N^{-1}\ln N + N_0^{-1}\right)$ with the estimates (26) and (27), which are the same as (9) and (10) for scheme (4) on piecewise-uniform grid (8), moreover, requirements imposed on the data of initial-boundary value problem (1) are the same for these schemes. However, unlike schemes on piecewise-uniform grids condensing in the layer region, in the solution decomposition method, the grid subproblems for regular and singular solution components are solved on *uniform grids* and the coefficients of the grid equations do not depend on the explicit form of the singular component of the solution.

Acknowledgements This research was supported by the Russian Foundation for Basic Research under grant No. 10-01-00726.

References

1. Bakhvalov, N.S.: On the optimization of methods for solving boundary value problems in the presence of a boundary layer. USSR Comput. Maths. Math. Phys. **9**. 139–166 (1969)
2. Doolan, E.P., Miller, J.J.H., Shilders, W.H.A.: Uniform Numerical Methods for Problem with Initial Boundary Layers. Boole Press, Dublin (1980)
3. Farrell, P.A., Hegarty, A.F., Miller, J.J.H., O'Riordan, E., Shishkin, G.I.: Robust Computational Techniques for Boundary Layers. Chapman and Hall/CRC, Boca Raton (2000)

4. Il'in, A.M.: Differencing scheme for a differential equation with a small parameter affecting the highest derivative. Math. Notes. **6**. 596–602 (1969)
5. Ladyzhenskaya, O.A., Solonnikov, V.A., Ural'ceva, N.N.: Linear and Quasilinear Equations of Parabolic Type. Amer. Math. Soc., Providence (1968)
6. Marchuk, G.I., Shaidurov, V.V.: Difference Methods and Their Interpolations. Springer–Verlag, New York Inc. (1983)
7. Miller, J.J.H., O'Riordan, E., Shishkin G.I.: Fitted numerical methods for singular perturbation problems. World Scientific, Singapore (1996)
8. Samarskii, A.A.: The Theory of Difference Schemes. Marcel Dekker, Inc., New York (2001)
9. Shishkin, G.I.: Approximation of solutions of singularly perturbed boundary value problems with a parabolic boundary layer. USSR Comput. Maths. Math. Phys. **29**. 1–10 (1989)
10. Shishkin, G.I.: Discrete Approximations of Singularly Perturbed Elliptic and Parabolic Equations. Ural Branch of Russian Academy of Sciences, Ekaterinburg (1992)
11. Shishkin, G.I., Shishkina, L.P.: Difference Methods for Singular Perturbation Problems. In: Chapman & Hall/CRC Monographs and Surveys in Pure and Applied Mathematics. **140**. CRC Press, Boca Raton (2009)
12. Shishkin, G.I., Shishkina, L.P.: Improved difference scheme of the solution decomposition method for a singularly perturbed reaction-diffusion equation. Trudy IMM UrO RAN. **16** 255–271 (2010). Translated in: Proceedings of the Steklov Institute of Mathematics. **272**. S197–S214 (2011)
13. Shishkin, G.I., Shishkina, L.P.: A Richardson Scheme of the Decomposition Method for Solving Singularly Perturbed Parabolic Reaction-Diffusion Equation. Comp. Math. Math. Phys. **50**. 2003–2022 (2010)
14. Shishkin, G.I., Shishkina, L.P.: Improved Approximations of the Solution and Derivatives to a Singularly Perturbed Reaction-Diffusion Equation Based on the Solution Decomposition Method. Comp. Math. Math. Phys. **51**. 1020–1049 (2011)

Implementation of the Continuous-Discontinuous Galerkin Finite Element Method

A. Cangiani, J. Chapman, E.H. Georgoulis, and M. Jensen

Abstract For the stationary advection-diffusion problem the standard continuous Galerkin method is unstable without some additional control on the mesh or method. The interior penalty discontinuous Galerkin method is more stable but at the expense of an increased number of degrees of freedom. The hybrid method proposed in [5] combines the computational complexity of the continuous method with the stability of the discontinuous method without a significant increase in degrees of freedom. We discuss the implementation of this method using the finite element library `deal.ii` and present some numerical experiments.

1 Introduction

We consider the advection-diffusion equation

$$-\varepsilon \Delta u + b \cdot \nabla u = f \qquad \text{in } \Omega \subset \mathbb{R}^d \tag{1}$$

$$u = g \qquad \text{on } \partial\Omega \tag{2}$$

with $0 < \varepsilon \ll 1$, $b \in W^{\infty}(\mathrm{div}, \Omega)$, $f \in L^2(\Omega)$ and $g \in H^{1/2}(\Omega)$. For simplicity we assume the region Ω is polygonal. We also assume $\rho := -\frac{1}{2}\nabla \cdot b \geq 0$ and then we have a weak solution $u \in H^1(\Omega)$.

J. Chapman (✉) · M. Jensen
Department of Mathematics, Durham University, Durham, UK
e-mail: john.chapman@durham.ac.uk; m.p.j.jensen@durham.ac.uk

A. Cangiani · E.H. Georgoulis
Department of Mathematics, University of Leicester, University Road, Leicester, UK
e-mail: andrea.cangiani@le.ac.uk; emmanuil.georgoulis@le.ac.uk

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_34,

It is well known that this problem can exhibit boundary or internal layers in the convection dominated regime and that for the standard continuous Galerkin (cG) formulation these layers cause non-physical oscillations in the numerical solution. Several adaptations to the cG method are effective but space does not allow their discussion here. We refer readers to [9] for a full description of these approaches. Discontinuous Galerkin (dG) methods also offer a more stable approach for approximating this problem. However the number of degrees of freedom required for dG methods is in general considerably larger than for cG methods.

We describe an alternative approach also studied in [5–7]. A dG method is applied on the layers and a cG method away from the layers. We call this approach the continuous-discontinuous Galerkin (cdG) method. The hypothesis is that provided the layers are entirely contained in the dG region the instability they cause will not propagate to the cG region. Note that in our formulation there are no transmission conditions at the interface between the two regions.

Here we present the cdG method and discuss its implementation using the `deal.ii` finite element library. We additionally provide some numerical experiments to highlight the performance of the method.

2 Finite Element Formulation

Assume that we can identify a decomposition of $\Omega := \Omega^{\mathrm{cG}} \cup \Omega^{\mathrm{dG}}$ where it is appropriate to apply the cG and dG methods respectively. We do not consider specific procedures to achieve this here, but generally it will be that we wish all boundary and internal layers to be within Ω^{dG}. Identifying these regions can be done a priori in some cases or a posteriori based on the solution of a dG finite element method. Consider a triangulation $\mathscr{T}_h$ of Ω which is split into two regions $\mathscr{T}_h^{\mathrm{cG}}$ and $\mathscr{T}_h^{\mathrm{dG}}$ where we will apply the cG and dG methods respectively. For simplicity we assume that the regions $\mathscr{T}_h^{\mathrm{cG}}$ and $\mathscr{T}_h^{\mathrm{dG}}$ are aligned with the regions Ω^{cG} and Ω^{dG} and the set J contains edges which lie in the intersection of the two regions. Call the mesh skeleton $\mathscr{E}_h$ and the internal skeleton $\mathscr{E}_h^o$. Define Γ as the union of boundary edges and the inflow and outflow boundaries by

$$\Gamma^{\mathrm{in}} = \{x \in \partial\Omega : b \cdot n \leq 0\}$$
$$\Gamma^{\mathrm{out}} = \{x \in \partial\Omega : b \cdot n > 0\}$$

where n is the outward pointing normal. Define Γ^{cG} (resp. Γ^{dG}) to be the intersection of Γ with $\mathscr{T}_h^{\mathrm{cG}}$ (resp. $\mathscr{T}_h^{\mathrm{dG}}$). By convention we say that the edges of J are part of the discontinuous skeleton $\mathscr{E}_h^{\mathrm{dG}}$ and $\mathscr{E}_h^{\mathrm{cG}} := \mathscr{E}_h \setminus \mathscr{E}_h^{\mathrm{dG}}$. With this convention there is potentially a discontinuity of the numerical solution at J. Elements of the mesh are denoted E, edges (resp. faces in 3d) by e and denote by h_E and h_e the diameter of an element and an edge, defined in the usual way.

The jump $[\![\cdot]\!]$ and average $\{\!\{\cdot\}\!\}$ of a scalar or vector function on the edges in $\mathscr{E}_h$ are defined as in, e.g., [1].

Definition 1. Define the cdG space to be

$$V_{\text{cdG}}^k := \{v \in L^2(\Omega) : v|_E \in \mathbb{P}^k, v|_{\partial\Omega\cap\partial\Omega^{\text{cG}}} = g, v|_{\Omega^{\text{cG}}} \in H^1(\Omega^{\text{cG}})\} \tag{3}$$

where $\mathbb{P}^k$ is the space of polynomials of degree at most k supported on E. This is equivalent to applying the usual cG space on Ω^{cG} and a dG space on Ω^{dG}.

We may now define the interior penalty cdG method: Find $u_h \in V_{\text{cdG}}^k$ such that for all $v_h \in V_{\text{cdG}}^k$

$$B_\varepsilon(u_h, v_h) = B_d(u_h, v_h) + B_a(u_h, v_h) = L_\varepsilon(f, g; v_h)$$

where

$$B_d(u_h, v_h) = \sum_{E\in\mathscr{T}_h}\left[\int_E \varepsilon\nabla u_h\cdot\nabla v_h - \int_E (b\cdot\nabla v_h)u_h - \int_E (\nabla\cdot b)u_h v_h\right]$$
$$+ \sum_{e\in\mathscr{E}_h}\left[\int_e \sigma\frac{\varepsilon}{h_e}[\![u_h]\!]\cdot[\![v_h]\!] - \int_e \left(\{\!\{\varepsilon\nabla u_h\}\!\}\cdot[\![v_h]\!] + \vartheta\{\!\{\varepsilon\nabla v_h\}\!\}\cdot[\![u_h]\!]\right)\right]$$

$$B_a(u_h, v_h) = \sum_{e\in\mathscr{E}_h^o}\int_e b\cdot[\![v_h]\!]u_h^- + \sum_{e\in\Gamma^{\text{out}}}\int_e (b\cdot n)u_h v_h$$

and

$$L_\varepsilon(f, g; v_h) = \sum_{E\in\mathscr{T}_h}\int_E f v_h + \sum_{e\in\Gamma}\left[\int_e\left(\sigma\frac{\varepsilon}{h_e}v_h - \vartheta\varepsilon\nabla\cdot v_h\right)g\right] - \sum_{e\in\Gamma^{\text{in}}}\int_e (b\cdot n)v_h g.$$

Here σ is the penalization parameter and $\vartheta \in \{-1, 0, 1\}$. Note that through the definition of V_{cdG}^k the edge terms are zero on $\mathscr{E}_h^{\text{cG}}$ and the method reduces to the standard cG FEM. If we take $\mathscr{T}_h = \mathscr{T}_h^{\text{dG}}$, i.e., the entire triangulation as discontinuous, we get the interior penalty (IP) family of dG FEMs (see, e.g., [1]).

The work of [8] shows that the cG method is the limit of the dG method as $\sigma \to \infty$. A reasonable hypothesis is that the solution to the cdG method is the limit of the solutions to the dG method as the penalty parameter $\sigma \to \infty$ on $e \in \mathscr{E}_h^{\text{cG}}$, i.e., super penalising the edges in $\mathscr{E}_h^{\text{cG}}$. Call σ_{cG} and σ_{dG} the penalty parameters for edges in $\mathscr{E}_h^{\text{cG}}$ and $\mathscr{E}_h^{\text{dG}}$ respectively. Call the numerical solution for the cdG problem $u_{\text{cdG},h} \in V_{\text{cdG}}^k$. The solution to the pure dG problem on the same mesh is denoted $u_{\text{dG},h} \in V_{\text{dG}}^k$ where V_{dG}^k is the usual piecewise discontinuous polynomial space on $\mathscr{T}_h$. Then we have:

Theorem 1. *The dG solution converges to the cdG solution of* (1)–(2) *as* $\sigma_{cG} \to \infty$, *i.e.,*

$$\lim_{\sigma_{cG}\to\infty}(u_{cdG,h} - u_{dG,h}) = 0.$$

We do not prove this result here but direct readers to [4] for a full discussion. Although this result does not imply stability of the cdG method (indeed, for the case where the Ω^{cG} region is taken to be the whole of Ω it shows that the cdG method has the same problems as the cG method), it does indicate that investigation of the cdG method as an intermediate stage between cG and dG is justified. Hence it aids in building an understanding of the convergence and stability properties of the cdG method, based on what is known for cG and dG. This, in turn, is of interest as the cdG method offers substantial reduction in the degrees of freedom of the method compared to dG.

3 Numerical Implementation

The cdG method poses several difficulties in implementation. One approach is to use the super penalty result of Theorem 1 to get a good approximation to the cdG solution. However this will give a method with the same number of degrees of freedom as dG. We therefore present an approach to implement the cdG method with the appropriate finite element structure. We discuss this approach with particular reference to the `deal.ii` finite element library [2, 3]. This is an open source C++ library designed to streamline the creation of finite element codes and give straightforward access to algorithms and data structures. We also present some numerical experiments.

3.1 Implementation in `deal.ii`

The main difficulty in implementing a cdG method in `deal.ii` is the understandable lack of a native cdG element type. In order to assign degrees of freedom to a mesh in `deal.ii` the code must be initialised with a `Triangulation` and then instructed to use a particular finite element basis to place the degrees of freedom. Although it is possible to initialise a `Triangulation` with the dG and cG regions set via the `material_id` flag, no appropriate element exists. In the existing `deal.ii` framework it would be difficult to code an element with the appropriate properties. A far more robust approach is to use the existing capabilities of the library and therefore allow access to other features of `deal.ii`. For instance without the correct distribution of degrees of freedom the resulting sparsity pattern of the finite element matrix would be suboptimal, i.e., containing more entries than required by the theory and therefore reducing the benefit of shrinking the number of degrees of freedom relative to a dG method.

The `deal.ii` library has the capability to handle problems with multiple equations applied to a single mesh such as the case of a elastic solid fluid interaction problem. In our case we wish to apply different methods to the same equation

on different regions of the mesh, which is conceptually the same problem in the `deal.ii` framework. In addition we will use the *hp* capability of the library.

The `deal.ii` library has the capability to create collections of finite elements, `hp::FECollection`. Here multiple finite elements are grouped into one data structure. As the syntax suggests the usual use is for *hp* refinement to create a set of finite elements of the same type (e.g., scalar Lagrange elements `FE_Q` or discontinuous elements `FE_DGQ`) of varying degree. Unfortunately it is not sufficient to create a `hp::FECollection` of cG and dG elements as the interface between the two regions will still be undefined. In order to create an admissible collection of finite elements we use `FE_NOTHING`. This is a finite element type in `deal.ii` with zero degrees of freedom. Using the `FESystem` class we create two vector-valued finite element types (`FE_Q`, `FE_NOTHING`) and (`FE_NOTHING`, `FE_DGQ`) and combine them in a `hp::FECollection`. We apply the first `FESystem` on the cG region, and the second on the dG region. Now when we create a `Triangulation` initialised with the location of cG and dG elements the degrees of freedom can be correctly distributed according to the finite element defined by `hp::FECollection`.

When assembling the matrix for the finite element method we need only be careful that we are using the correct element of `hp::FECollection` and the correct part of `FESystem`. The most difficult case is on the boundary J where from a dG element we must evaluate the contribution from the neighbouring cG element (recall that in the cdG method a jump is permissible on J).

If we implement the cdG method in `deal.ii` in this way we create two solutions: one for the `FE_Q`-`FE_NOTHING` component and another for the `FE_NOTHING`-`FE_DGQ` component. Consider a domain $\Omega = (0,1)^2$ in $\mathbb{R}^2$, $b = (1,1)^\top$ and $\vartheta = -1$. The Dirichlet boundary conditions and the forcing function f are chosen so that the analytical solution is

$$u(x,y) = x + y(1-x) + \frac{\mathrm{e}^{-\frac{1}{\varepsilon}} - \mathrm{e}^{-\frac{(1-x)(1-y)}{\varepsilon}}}{1 - \mathrm{e}^{-\frac{1}{\varepsilon}}}. \tag{4}$$

This solution exhibits an exponential boundary layer along $x = 1$ and $y = 1$ of width $\mathscr{O}(\varepsilon)$. For illustration we solve the finite element problem on a 256 element grid and fix $\Omega^{\mathrm{cG}} = [0, 0.6875)^2$, $\varepsilon = 10^{-6}$ and show the results in Fig. 1. This Ω^{cG} region larger than is required for stability (see Example 1 below) but shows the behaviour of `FE_NOTHING` more clearly. We show each of the components of `FE_SYSTEM` and the combined solution. For comparison we also show the dG finite element solution for the same problem.

One advantage of following the `deal.ii` framework is that the data structures will allow the implementation of *hp* methods. In fact we can envisage the implementation of a "*hpe* method" in which refinement may be undertaken in either mesh size h, polynomial degree p or element type e (either cG or dG elements). We propose no specific scheme here but simply remark that implementing a *hpe* method is relatively straightforward with the `FE_NOTHING` approach.

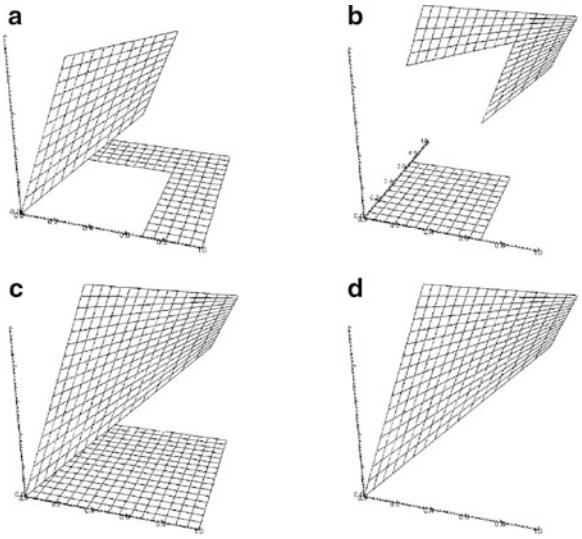


Fig. 1 Solution components of `FE_NOTHING` implementation applied to Example 1 with $\varepsilon = 10^{-6}$. (**a**) cG-`FE_NOTHING`. (**b**) `FE_NOTHING`-dG. (**c**) Combined cdG solution. (**d**) dG solution

3.2 Numerical Examples

We present two numerical experiments highlighting the performance of the cdG method. Both examples present layers when ε is small enough. In each case we fix the region where the continuous method is to be applied then vary ε. This causes the layer to steepen. In the advection dominated regime, i.e., ε large and no steep layer present, we see the cdG solution approximates the true solution well. As we make ε smaller the layer forms and extends into the continuous region. As ε becomes smaller still the layer leaves the continuous region and the performance of the dG and cdG method is indistinguishable. In each experiment we pick the Ω^{cG} and Ω^{dG} regions so that with the given refinement the region $\mathscr{T}_h^{\mathrm{dG}}$ consists of exactly one layer of elements and coincides with Ω^{dG}.

We use bilinear elements on quadrilaterals. In each example the time for matrix assembly and solution (using a direct solver) is faster by a factor of approximately 3 using the cdG method as opposed to the dG method. The dG method in either example has 4,096 degrees of freedom compared to 1,276 (resp. 1,312) for the cdG method in Example 1 (resp. Example 2).

Example 1. Consider again the problem with true solution (4) presented above. We solve the finite element problem on a 1,024 element grid and fix $\Omega^{\mathrm{cG}} = [0, 0.96875)^2$ so exactly one row of elements is in Ω^{dG}. As we vary ε the layer sharpens and moves entirely into the dG region.

As we can see from Fig. 2 before the layer has formed the two methods perform well. As the layer begins to form with decreasing ε it is not entirely contained in the discontinuous region and the error peaks. As the layer sharpens further it is entirely

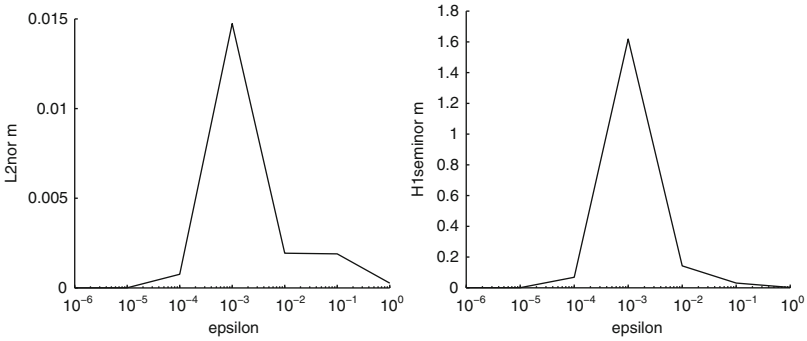


Fig. 2 Decreasing ε with a fixed Ω decomposition in Example 1. The maximum difference in either norm occurs when the layer is sharp but not contained entirely in Ω^{dG}

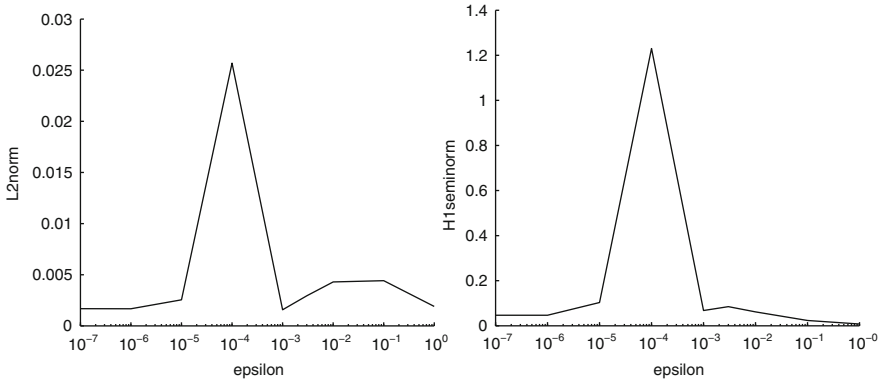


Fig. 3 Decreasing ε with a fixed Ω decomposition in Example 2. As in Fig. 2 the maximum difference in either norm occurs when the layer is sharp but not contained entirely in Ω^{dG}

contained in the discontinuous region and the difference between the two solutions becomes negligible.

Example 2. Now we look at a problem with an internal layer. Let the advection coefficient be given by $b = (-x, y)^{\top}$ and pick the boundary conditions and right hand side f so that the true solution is

$$u(x, y) = (1 - y^2)\mathrm{erf}\left(\frac{x}{\sqrt{2\varepsilon}}\right),$$

where erf denotes the error function.

We solve on the region $\Omega = (-1, 1)^2$. The solution has an internal layer along $y = 0$ of width $\mathcal{O}(\sqrt{\varepsilon})$ and we fix $\Omega^{\mathrm{cG}} = \{(x, y) : x \in [-1, -0.0625) \cup (0.0625, 1], y \in [-1, 1]\}$.

In Fig. 3 we notice the same behaviour as in Example 1. When the internal layer is present it must be contained within the discontinuous region for the two methods

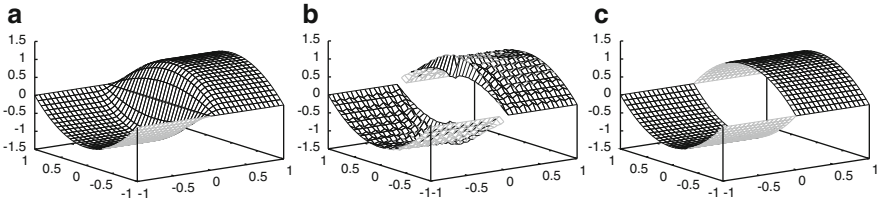


Fig. 4 The cdG solutions for Example 2 for various ε. When $\varepsilon = 10^{-4}$ the layer is steep enough to cause oscillations but not sharp enough to be contained entirely in Ω^{dG}. In this case the oscillations are clearly visible, but they are not present when $\varepsilon = 10^{-6}$ as the layer has moved entirely within Ω^{dG}. (**a**) $\varepsilon = 10^{-2}$. (**b**) $\varepsilon = 10^{-4}$. (**c**) $\varepsilon = 10^{-6}$

to perform equivalently. In Fig. 4 we can see the cdG solution for various ε with the oscillations clearly visible when $\varepsilon = 10^{-4}$. When the layer is sharpened, the oscillations disappear.

References

1. Arnold, D.N., Brezzi, F., Cockburn, B., Marini, L.D.: Unified analysis of discontinuous Galerkin methods for elliptic problems. SIAM J. Numer. Anal. **39**(5), 1749–1779 (2001)
2. Bangerth, W., Hartmann, R., Kanschat, G.: deal.II – a general purpose object oriented finite element library. ACM Trans. Math. Softw. **33**(4), 24/1–24/27 (2007)
3. Bangerth, W., Kanschat, G.: `deal.II` Differential Equations Analysis Library, Technical Reference. http://www.dealii.org
4. Cangiani, A., Chapman, J., Georgoulis, E.H., Jensen, M.: Super penalties for the continuous discontinuous Galerkin method. In Preparation
5. Cangiani, A., Georgoulis, E.H., Jensen, M.: Continuous and discontinuous finite element methods for convection-diffusion problems: A comparison. In: International Conference on Boundary and Interior Layers. Göttingen (2006)
6. Dawson, C., Proft, J.: Coupling of continuous and discontinuous Galerkin methods for transport problems. Comput. Meth. in Appl. Mech. and Eng. **191**(29–30), 3213 – 3231 (2002)
7. Devloo, P.R.B., Forti, T., Gomes, S.M.: A combined continuous-discontinuous finite element method for convection-diffusion problems. Lat. Am. J. Solids Stru. **2**(3), 229–246 (2007)
8. Larson, M.G., Niklasson, A.J.: Conservation properties for the continuous and discontinuous Galerkin methods. Tech. Rep. 2000–08, Chalmers University of Technology (2000)
9. Roos, H.G., Stynes, M., Tobiska, L.: Numerical Methods for Singularly Perturbed Differential Equations: Convection-Diffusion and Flow Problems, Second edn. Springer-Verlag, Berlin (2008)

Towards A Posteriori Error Estimators for Realistic Problems in Incompressible Miscible Displacement

J. Chapman and M. Jensen

Abstract The incompressible miscible displacement problem has attracted interest in recent years as it models economically important activities such as oil recovery and groundwater flow. It is important that numerical simulations can accurately model the types of problems seen in industry. We discuss a posteriori finite element indicators for the incompressible miscible displacement problem and propose an extension to a mixed-discontinuous Galerkin scheme. Furthermore we highlight some physically realistic scenarios not covered by the existing analysis and outline the theory of weighted spaces required to address them.

1 Introduction

We consider the problem of finding the numerical solution to the coupled equations for the pressure $p = p(t,x)$, Darcy velocity $u = u(t,x)$ and concentration $c = c(t,x)$ of one incompressible fluid in a porous medium being displaced by another. We consider the miscible case and where both fluids are in the same phase.

Consider the domain $\Omega_T := J \times \Omega$, where $J := (0,T]$ is the time interval and Ω is a bounded Lipschitz polygonal domain in $\mathbb{R}^d$ so that we may construct a straight edged polygonal mesh that fits the domain exactly. The equations for the miscible displacement of one incompressible fluid by another in a porous medium on Ω_T are given by

J. Chapman (✉) · M. Jensen
Durham University, Durham, UK
e-mail: john.chapman@durham.ac.uk; m.p.j.jensen@durham.ac.uk

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_35, © Springer-Verlag Berlin Heidelberg 2013

$$\varphi \frac{\partial c}{\partial t} + u \cdot \nabla c - \nabla \cdot (\mathbb{D}(u) \nabla c) + c q^I = \hat{c} q^I \tag{1}$$

$$u = -\frac{\mathbb{K}}{\mu(c)} \left(\nabla \cdot p - \rho(c) g \right) \tag{2}$$

$$\nabla \cdot u = q^I - q^P, \tag{3}$$

with the boundary conditions on $\partial \Omega_T := J \times \partial \Omega$

$$u \cdot n = 0 \tag{4}$$

$$(\mathbb{D}(u) \nabla c) \cdot n = 0 \tag{5}$$

and the initial conditions

$$c(0, \cdot) = c_0. \tag{6}$$

We denote by: $\varphi(x)$ the porosity of the medium; $q^I \geq 0$ and $q^P \geq 0$ the pressure at injected (source) and production (sink) wells; $\mathbb{K}(x)$ the absolute permeability of the medium; $\mu(c)$ the viscosity of the fluid mixture; $\rho(c)$ the density of the fluid mixture; g the constant vector of gravity; $\mathbb{D}(u, x)$ the diffusion-dispersion tensor; $\hat{c}$ the injected concentration; c_0 the initial concentration; and n the unit outward normal vector on $\partial \Omega$. We will define $a^{-1}(c) := \mathbb{K}^{-1} \mu$. The coupling is non-linear through the coefficients $\mathbb{D}(u, x)$, $\mu(c)$ and the advection term.

This model to describe incompressible miscible displacement has several industrial applications including enhanced oil recovery (EOR) and groundwater flow [4, 15, 17]. In both of these examples an injected fluid (carbon dioxide resp. contaminated water) mixes with a fluid in a reservoir of porous rock filled with a second fluid (oil resp. fresh water). The goal of a numerical simulation is to accurately predict the spread of the injected fluid through the medium, for instance to model the spread of contaminants. In practice this is considerably complicated by the presence of irregularly shaped reservoirs, cracks and highly heterogeneous media. However existing a posteriori error estimators rely on estimates which do not apply in some reasonable cases. Consider for example the case of an L-shaped domain in $\mathbb{R}^2$ as studied in [2, Numerical Example 2] and shown in Fig. 1. For injection and production wells located at $(1, 1)$ and $(0, 0)$ respectively, and with discontinuous permeability, we see a singular velocity field at $(0.5, 0.5)$, the location of the re-entrant corner. This demonstrates the limitations of the analysis of existing a posteriori papers and presents a challenge when applying numerical simulations to real world problems.

In this proceeding we will review the existing literature for the miscible displacement problem in Sect. 2, including discussion of a new result for a discontinuous-mixed estimator in the existing framework. In Sect. 3 we will consider deficiencies in the existing literature with regard to problems with singular coefficients and present a brief introduction to a proposed method involving weighted spaces.

2 A Posteriori Estimators on Simple Domains

The derivation of equations for this problem can be found in many texts, e.g., [3]. The existence of solutions to (1)–(6) under physically realistic assumptions was shown by Chen and Ewing, [7]. The numerical simulation of this problem has been widely studied and we provide a brief review before focussing more closely on a posteriori estimators.

We direct the reader to the book by Peaceman, [20] for a review of early finite difference approaches. For the finite element method the first appearance in the literature is the work of Ewing and Wheeler, [13] who developed an a priori estimator for the pressure and concentration components. A scheme using a mixed-continuous Galerkin (cG) method to solve for pressure, velocity and concentration independently was introduced by Douglas et al., [10] and the pressure and velocity components were further studied by Douglas and Roberts, [11].

The mixed-discontinuous Galerkin (dG) method was presented by Sun et al., [22] where cut off functionals were used to ensure that the velocity is bounded and to make the combined scheme converge. Other papers of interest include work by Cui, [8,9] covering the compressible case and by Larson and Målqvist, [16] looking at the dual weighted residual method in the case of one way coupling. The case of low regularity has been studied by Bartels et al., [2] and Rivière and Walkington, [21].

Of particular interest here is the work of Chen and Liu, [6] who have studied a posteriori error estimates for the mixed-cG method and Yang, [23] who has considered a mixed-dG method for the compressible problem. We will review the work of [6] in more detail and briefly describe our extension to a dG method which differs substantially from that of [23].

In [6] the authors consider a domain $\Omega \subset \mathbb{R}^2$ and use the spaces

$$V := H(\mathrm{div}; \Omega) = \{v \in (L^2(\Omega))^2 : \nabla \cdot v \in L^2(\Omega), v \cdot n = 0 \text{ in } H^{-1/2}(\partial\Omega)\}$$

and

$$W := \{w \in L^2(\Omega) : \int_\Omega w \, \mathrm{d}x = 0\}.$$

Some method must be employed to approach the non-linear system, so an auxiliary equation is introduced using the numerical concentration c_h in place of the concentration c, namely: Find $(\tilde{u}, \tilde{p}) \in V \times W$ such that

$$(a^{-1}(c_h)\tilde{u}, v) - (\nabla \cdot v, \tilde{p}) = (\rho(c_h)g, v) \qquad \forall v \in V, t \in J \tag{7}$$

$$(\nabla \cdot \tilde{u}, w) = (q^I - q^P, w) \qquad \forall w \in W, t \in J. \tag{8}$$

Then the authors use the stability of the mixed weak form to produce the following bound in the manner of, e.g., [10]

$$\|u - \tilde{u}\|_V + \|p - \tilde{p}\|_W \leq K\|\tilde{u}\|_{L^\infty(\Omega)}\|\mathrm{E_c}\|_{L^2(\Omega)} \tag{9}$$

where $\mathrm{E_c} := c - c_h$. This result forms half of the analysis for the pressure and velocity terms, the second part being a duality argument for the auxiliary problem

$$\nabla\cdot\xi = \tilde{p} - p_h \qquad \text{on } \Omega_T \tag{10}$$

$$\xi = -a(c_h)\nabla\psi \qquad \text{on } \Omega_T \tag{11}$$

$$\xi\cdot n = 0 \qquad \text{on } \partial\Omega_T, \tag{12}$$

which is assumed to have elliptic regularity

$$\|\psi\|_{H^2(\Omega)} \leq C\|p - \tilde{p}\|_{L^2(\Omega)}. \tag{13}$$

Using a saturation assumption it is then possible to bound $\|\tilde{p} - p_h\|_W$ and $\|\tilde{u} - u_h\|_V$ which can then be combined with (9) to bound $\mathrm{E_p} := p - p_h$ and $\mathrm{E_u} := u - u_h$ with a posteriori terms and $\|\mathrm{E_c}\|_{L^2(\Omega)}$.

For the concentration terms a backward parabolic equation is used:

$$\varphi\frac{\partial\zeta}{\partial t} + \nabla\cdot(u_h\zeta) + \nabla\cdot(\mathbb{D}(u_h)\nabla\zeta) - q'\zeta = \tilde{c} - c_h \qquad \text{on } \Omega_T \tag{14}$$

$$\big(\mathbb{D}(u_h)\nabla\zeta\big)\cdot n = 0 \qquad \text{on } \partial\Omega_T \tag{15}$$

$$\zeta(T, x) = 0 \qquad \text{for } x \in \Omega. \tag{16}$$

Using the regularity results of, e.g., [12, Chap. 7] the authors have

$$\operatorname*{ess\,sup}_{t\in J}\|\zeta\|_{H^1(\Omega)} + \|\zeta\|_{L^2(J;H^2(\Omega))} \leq C\|\tilde{c} - c_h\|_{L^2(J;L^2(\Omega))}. \tag{17}$$

With this an a posteriori bound on $\|\tilde{c} - c_h\|_{L^2(\Omega)}$ can be formulated and then using Gronwall's Lemma a bound for $\|\mathrm{E_c}\|_{L^2(\Omega)}$ can be found. This is combined with the pressure-velocity bound to complete the paper.

2.1 Extension to a Discontinuous Galerkin Scheme

The extension of the scheme of [6] to discontinuous finite elements is non-trivial in several respects. We will not cover the full theory here but highlight main points for a valid dG-mixed scheme.

All of the fundamental problems come from the discontinuous space used to approximate the concentration. In (7) and (11) we use the standard L^2 projection on the c_h terms so we have sufficient regularity to produce an estimate as in (9) and guarantee elliptic regularity as in (13). The analysis then proceeds on this region as before.

For the concentration analysis we use the true velocity u in (14)–(16) and we replace $\tilde{c} - c_h$ by $\mathtt{E}_c$ on the right hand side of (14). This is necessary as applying integration by parts cell wise now results in jump terms across the inter-element boundaries and using the continuous u allows cancellation of difficult terms. We do not require a saturation assumption. However problematic terms still remain and a Sobolev embedding and careful splitting of the terms is required to be able to control the jumps of u_h on the edges. Finally the result on each region is combined to give an a posteriori estimator.

A complete analysis of this problem will be presented in [5].

3 Extensions to Realistic Domains

The extension to a discontinuous scheme is of technical interest but some assumptions used are not appropriate to practical settings. We discuss these assumptions here and briefly present theoretical results which can be used to relax them.

The primary difficulty in extending this work toward industrially realistic problems is the formation of points with unbounded velocity. As discussed in the introduction cases with this feature are not difficult to formulate, for example the case of an L-shaped domain. In Fig. 1 we see a representation of this on an L-shaped domain with decreased permeability on $(0, 0.5)^2$. The velocity across the re-entrant corner is unbounded. The figure is in fact a numerical approximation to this problem using the implementation of [14]. We approximate the injection and extraction points over a single element although it is recognized that true point sources can cause difficulties, see for example [10, Sect. 8].

The boundedness of the velocity is assumed implicitly in (9). We also see that an unbounded velocity would not admit the usual regularity bounds for (10) and (11) and (14) which are a fundamental part of the analysis. We restrict ourselves to the case of point singularities in the coefficients of the problem and show that regularity results and bounds similar to (9), (13) and (17) exist in weighted spaces and can be used to formulate a new a posteriori estimator.

3.1 Weighted Spaces

We construct a set of vertices $Q_i \in \mathscr{V}$ where the solution of (1) and (6) is singular. We define the weighted Sobolev space (Babuška-Kondratiev space) by

$$\mathscr{K}_a^m(\Omega) = \{f : \vartheta^{|\alpha|-a} D^\alpha f \in L^2(\Omega), \forall |\alpha| \leq m\}$$

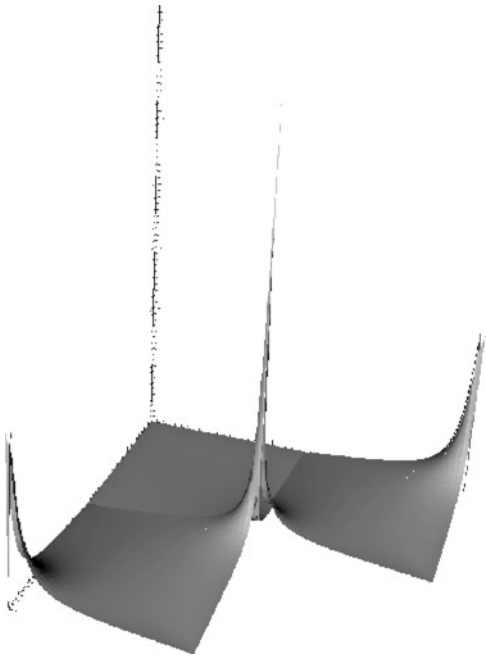


Fig. 1 A numerical simulation showing an approximation to the unbounded velocity field on an L-shaped domain with a re-entrant corner at $(0.5, 0.5)$ and singular injection/extraction points at $(0, 0)$ and $(1, 1)$ respectively. The tendency to unboundedness of the velocity at the injection/extraction points and corner is clearly apparent

with $\vartheta = \Pi_i r_i$ and r_i the distance from a point $x \in \Omega$ to a vertex Q_i. It is possible to show a smoothed version of ϑ exists and has the same properties (see [18, Sect. 2.3.1]) and we will henceforth used this smoothed version. Define the norm on this space to be

$$\|v\|^2_{\mathscr{K}^m_a(\Omega)} = \sum_{|\alpha|\leq m} \|\vartheta^{|\alpha|-a} D^\alpha v\|^2_{L^2(\Omega)}$$

and the inner product

$$(u, v)_{\mathscr{K}^m_a(\Omega)} = \sum_{|\alpha|\leq m} \int_\Omega \vartheta^{2(|\alpha|-a)} (D^\alpha u)(D^\alpha v).$$

The spaces $\mathscr{K}^m_a(\partial\Omega)$ on the boundary and semi-norms are defined analogously.

Using [19, Theorem 2.3] and adjusting notation we have for Ω partitioned into a domain $\mathscr{T}_h$ the following result.

Corollary 1. *For $m \geq 1$ and provided that $(\tilde{p} - p_h)|_E \in \mathscr{K}^{m-1}_{a-1}(E)$ the solution ψ of (10)–(11) is in $\mathscr{K}^1_{a+1}(\Omega)$ and we have the estimate*

$$\|\psi\|_{\mathscr{K}^2_{a+1}(E)} \leq C \sum_{E\in\mathscr{T}_h} \|\tilde{p} - p_h\|_{\mathscr{K}^0_{a-1}(E)} \tag{18}$$

for some constant $C > 0$.

Remark 1. In the limit case $a = 1$ we see that this result is stronger than that used in [6]. In this case we find

$$\|\psi\|^2_{\mathscr{K}^2_2(E)} = |\psi|^2_{H^2(E)} + |\psi|^2_{\mathscr{K}^1_1(E)} + \|\psi\|^2_{\mathscr{K}^0_2(E)} \leq \left(\sum_{E \in \mathscr{T}_h} \|\tilde{p} - p_h\|_{L^2(E)} \right)^2$$

and we note that spaces with positive lower exponent involve dividing by the weight so they are larger than the equivalent L^2 or H^1 (semi) norm.

We require also an interpolation inequality on the weighted spaces. This is an extension to the results in [19].

Lemma 1. *Provided a is chosen such that* $\psi \in \mathscr{K}^{m+1}_{a+1}(E)$ *and the mesh is divided isotropically there exists a constant* $C > 0$ *depending on m but not on* h_E *such that*

$$\|\psi - \psi_I\|_{\mathscr{K}^{m'}_1(E)} \leq C h_E^{m+1-m'} \|u\|_{\mathscr{K}^{m+1}_{a+1}(E)} \quad \forall \psi \in \mathscr{K}^{m+1}_{a+1}(E), m+1 \geq m' \geq 1$$

where ψ_I *is the degree m Lagrange interpolant of* ψ.

The final result needed to solve the pressure-velocity problem is an analogue of (9) for which we need slightly more regularity.

Theorem 1. *Let* (p, u) *and* $(\tilde{p}, \tilde{u})$ *be defined as in* (2), (3) *and* (7), (8) *respectively. Then for* $\tilde{p} \in \mathscr{K}^2_{a+1}(\Omega)$ *we have that* $(\vartheta^{a-1}\tilde{u}) \in L^\infty(\Omega)$ *and for some constant* $C > 0$

$$\|u - \tilde{u}\|_V + \|p - \tilde{p}\|_W \leq C \|\vartheta^{a-1}\tilde{u}\|_{L^\infty(\Omega)} \|\mathtt{E}_\mathtt{C}\|_{\mathscr{K}^0_{1-a}(\Omega)}.$$

Proof. Via the definitions of the weighted spaces and a Sobolev embedding we can show that $\vartheta^{a-1}\tilde{u}$ is bounded. Then following the steps of [10] but multiplying and dividing by ϑ^{a-1} on the right hand side we complete the proof. □

With these results the construction of an a posteriori bound on the error in the velocity and pressure follows as in the discontinuous analysis with careful consideration of the weighted spaces when integrating.

The analysis for the concentration equation is more difficult. As we use the velocity in (14) the first and second order coefficients are unbounded. The result (18) extends to the steady state case, i.e., for $\frac{\partial \zeta}{\partial t} = 0$. However it is not clear that the elliptic result in Corollary 1 will extend to a regularity result of the type (17) for the parabolic equation (14). However if we make the assumption that we can show such a result the a posteriori analysis for the transport equation is possible and gives a bound on $\mathtt{E}_\mathtt{C}$ in weighted norms of known numerical terms.

For further technical results on these weighted spaces we direct readers to [1] and the references therein in addition to the citations already given.

4 Conclusions and Future Work

Following a review of previous work in the field we have outlined the steps required to extend the a posteriori results of [6] to a mixed-dG scheme on both simple domains and also on those domains resulting in unbounded velocity. The extension of the elliptic regularity result in Corollary 1 to the parabolic case remains an open problem.

Our future work on this problem will be give complete details of the mixed-dG method including comprehensive numerical experiments. Additionally we plan to publish a paper giving full details of the mixed-dG scheme using weighted spaces and a comparison of the weighted and unweighted estimators for problems of both bounded and unbounded velocity.

References

1. Ammann, B., Nistor, V.: Weighted Sobolev spaces and regularity for polyhedral domains. Comp. Meth. in Appl. Mech. and Eng. **196**(37–40), 3650–3659 (2007)
2. Bartels, S., Jensen, M., Müller, R.: Discontinuous Galerkin finite element convergence for incompressible miscible displacement problems of low regularity. SIAM J. Numer. Anal. **47**(5), 3720–3743 (2009)
3. Bear, J.: Dynamics of Fluids in Porous Media. Dover, New York (1988)
4. Bear, J., Cheng, A.H.D.: Modeling Groundwater Flow and Contaminant Transport. Theory and Applications of Transport in Porous Media. Springer (2009)
5. Chapman, J.: Finite element methods for the incompressible miscible displacement problem. Ph.D. Thesis, University of Durham (2012). In preparation
6. Chen, Y., Liu, W.: A posteriori error estimates of mixed methods for miscible displacement problems. Int. J. Numer. Meth. Eng. **73**(3), 331–343 (2008)
7. Chen, Z., Ewing, R.: Mathematical analysis for reservoir models. SIAM J. Math. Anal. **30**(2), 431–453 (1999)
8. Cui, M.: A combined mixed and discontinuous Galerkin method for compressible miscible displacement problem in porous media. J. Comput. Appl. Math. **198**(1), 19–34 (2007)
9. Cui, M.: Analysis of a semidiscrete discontinuous Galerkin scheme for compressible miscible displacement problem. J. Comput. Appl. Math. **214**(2), 617–636 (2008)
10. Douglas, J., Ewing, R.E., Wheeler, M.F.: The approximation of the pressure by a mixed method in the simulation of miscible displacement. RAIRO Anal. Numer. **17**, 17–33 (1983)
11. Douglas, J., Roberts., J.E.: Global estimates for mixed methods for second order elliptic equations. Math. Comp. **44**(169), 39–52 (1985)
12. Evans, L.C.: Partial Differential Equations. American Mathematical Society, Rhode Island (2008)
13. Ewing, R.E., Wheeler, M.F.: Galerkin methods for miscible displacement problems in porous media. SIAM J. Numer. Anal. **17**(3), 351–365 (1980)
14. Jensen, M., Müller, R.: Stable Crank-Nicolson discretisation for incompressible miscible displacement problems of low regularity. In: G. Kreiss, P. Ltstedt, A. Mlqvist, M. Neytcheva (eds.) Numerical Mathematics and Advanced Applications 2009, pp. 469–477. Springer Berlin Heidelberg (2010)
15. Lake, L.W.: Enhanced Oil Recovery. Prentice Hall (1996)
16. Larson, M.G., Målqvist, A.: Goal oriented adaptivity for coupled flow and transport problems with applications in oil reservoir simulations. Comp. Meth. Appl. Mech. Eng. **196**(37–40), 3546–3561 (2007)

17. Latil, M.: Enhanced Oil Recovery. Institut Français du Pétrole Publications. Éditions Technip (1980)
18. Li, H.: Elliptic equations with singularities: A priori analysis and numerical approaches. PhD Thesis, The Pennsylvania State University (2008)
19. Li, H., Mazzucato, A., Nistor, V.: Analysis of the finite element method for transmission/mixed boundary value problems on general polygonal domains. Elec. Trans. Numer. Anal. **37**, 41–69 (2010)
20. Peaceman, D.W.: Fundementals of Numerical Reservoir Simulation. Developments in Petroleum Science. Elsevier Scientific Publishing Company (1977)
21. Rivière, B., Walkington, N.: Convergence of a discontinuous Galerkin method for the miscible displacement equation under low regularity. SIAM J. Numerical Analysis **49**(3), 1085–1110 (2011)
22. Sun, S., Rivière, B., Wheeler, M.F.: A combined mixed finite element and discontinuous Galerkin method for miscible displacement problem in porous media. In: Recent Progress in Computational and Applied PDEs, pp. 323–351. Kluwer/Plenum, New York (2002)
23. Yang, J.: A posteriori error of a discontinuous Galerkin scheme for compressible miscible displacement problems with molecular diffusion and dispersion. Int. J. Numer. Meth. Fluids **65**(7), 781–797 (2009)

Application of hp–Adaptive Discontinuous Galerkin Methods to Bifurcation Phenomena in Pipe Flows

K.A. Cliffe, E.J.C. Hall, and P. Houston

Abstract In this article we consider the a posteriori error estimation and adaptive mesh refinement of hp–version discontinuous Galerkin finite element approximations of the bifurcation problem associated with the steady incompressible Navier–Stokes equations. Particular attention is given to the reliable error estimation of the critical Reynolds number at which a steady pitchfork bifurcation occurs when the underlying physical system possesses rotational and reflectional or $O(2)$ symmetry. Here, computable a posteriori error bounds are derived based on employing the generalization of the standard Dual Weighted Residual approach, originally developed for the estimation of target functionals of the solution, to bifurcation problems. Numerical experiments highlighting the practical performance of the proposed a posteriori error indicator on hp–adaptively refined computational meshes are presented.

1 Introduction

In this article we study the stability of the three–dimensional incompressible Navier-Stokes equations in the case when the underlying system possesses both rotational and reflectional symmetry, or more precisely, $O(2)$ symmetry. To this end, we are interested in numerically estimating the critical Reynolds number Re^c, at which a (pitchfork) bifurcation point first occurs; a review of techniques for bifurcation detection can be found in Cliffe et al. [8], for example. The detection of bifurcation points in this setting is now well understood, for example, see Golubitsky and Schaeffer [11]. For the purposes of this article, we assume that a symmetric steady

K.A. Cliffe · E.J.C. Hall · P. Houston (✉)
School of Mathematical Sciences, University of Nottingham, University Park, Nottingham NG7 2RD, UK
e-mail: Andrew.Cliffe@nottingham.ac.uk; Edward.Hall@nottingham.ac.uk; Paul.Houston@nottingham.ac.uk

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_36,

state solution to the incompressible Navier-Stokes equations undergoes a steady pitchfork bifurcation at a critical value of the Reynolds number. Estimation of the critical Re^c can be undertaken by discretizing a suitable extended system of partial differential equations; see Brezzi et al. [3] and Werner and Spence [14] for steady bifurcations. For discretization purposes we exploit the hp–version of the interior penalty discontinuous Galerkin (DG) method [9, 10]. The derivation of a computable error estimator for the critical parameter of interest, namely Re^c, based on exploiting the Dual Weighted Residual (DWR) a posteriori error estimation technique is undertaken and implemented within an automatic hp–adaptive finite element algorithm. The work presented in this article extends our recent work in [5–7] based on employing standard mesh subdivision adaptive algorithms (h–refinement) to the hp–setting. The performance of the proposed hp–refinement strategy will be demonstrated for the benchmark problem of determining the symmetry breaking bifurcation of an incompressible flow in a cylindrical pipe with a stenotic region.

2 Detecting Steady Bifurcation Points in the Presence of $O(2)$ Symmetry

Suppose we have a nonlinear, time dependent problem of the form

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

where F is a map from $V \times \mathbb{R} \to V$, for some Banach space V, with norm $\|\cdot\|$. Here, λ is some distinguished parameter and u is the state variable. For the purposes of this article, F will represent the incompressible Navier–Stokes equations, written in cylindrical coordinates, in a generic open cylindrical pipe geometry $\Omega \subset \mathbb{R}^3$, subject to appropriate boundary conditions, with V a subset of $H^1(\Omega)^3 \times L_2(\Omega)$. With this in mind, λ is identified as the Reynolds number.

We denote the Fréchet derivative of F with respect to u at a fixed point $(w, \chi) \in V \times \mathbb{R}$ by $F'_u(w, \chi; \cdot)$ and similarly the derivative with respect to λ by $F'_\lambda(w, \chi)$. Here and throughout this article, we use the convention that in semi-linear forms such as $F'_u(\cdot, \cdot; \cdot)$ the form is linear with respect to all arguments to the right of the semicolon. Higher order Fréchet derivatives are expressed in much the same manner, cf. [7], for example. We assume that $F'_u(u, \lambda; \cdot) : V \to V$ is Fredholm of index 0 for all $(u, \lambda) \in V \times \mathbb{R}$. With this notation, a steady bifurcation of (1) satisfies the following extended system: find $\mathbf{u} := (u, \phi, \lambda)$ such that

$$T(\mathbf{u}) \equiv \begin{pmatrix} F(u, \lambda) \\ F'_u(u, \lambda; \phi) \\ \langle \phi, g \rangle - 1 \end{pmatrix} = \mathbf{0}, \tag{2}$$

where $\langle \cdot, \cdot \rangle$ denotes the duality pairing between the spaces V and V', V' being the dual space of V, and $g \in V'$ is some suitable functional satisfying $\langle \phi, g \rangle \neq 0$, cf. [7]. The equation $\langle \phi, g \rangle - 1 = 0$ acts to normalise the null-function ϕ, thus ensuring that, if a solution to (2) exists at some λ, the solution is unique.

We now discuss how to reduce the complexity of (2) when the problem at hand possesses some underlying symmetry, in our case $O(2)$ symmetry. We recall that $O(2)$ is the Lie group which comprises rotations and reflections. With this in mind, we assume that F is $O(2)$ equivariant, cf. [7, 8], and write $V^{O(2)}$ to denote the subspace of V consisting of all $O(2)$ invariant functions.

It is a standard result, see Aston [1], that, for the $O(2)$ case, there exists a unique orthogonal decomposition of V, namely,

$$V = \sum_{m=0}^{\infty} V^m, \quad V^m \perp V^l, \quad m \neq l, \tag{3}$$

where each V^m is an $O(2)$ invariant subspace of V, with the property that V^m is irreducible, i.e., V^m has no proper $O(2)$ invariant subspaces. The decomposition (3) is produced purely by the rotation elements of $O(2)$ and the fact that there is also a reflection element has not been used. Indeed, there is a finer decomposition

$$V^m = V^{m,s} \oplus V^{m,a}, \quad m = 1, 2, \ldots,$$

where $V^{m,s}$ and $V^{m,a}$ are the symmetric and anti–symmetric components of V^m, respectively. Further, $V^{m,s}$ and $V^{m,a}$ are invariant subspaces of V with respect to the reflection.

With this notation, it can be shown (see [7, 8] for details) that the critical parameter values at which symmetric steady state solutions $u \in V^{O(2)}$ to (1) lose stability may be located by solving the following problems for $m = 0, 1, \ldots$: find $\mathbf{u} = (u, \phi, \lambda) \in V^{O(2)} \times V^{m,s} \times \mathbb{R}$ (or equivalently $\mathbf{u} = (u, \phi, \lambda) \in V^{O(2)} \times V^{m,a} \times \mathbb{R}$) such that

$$T(\mathbf{u}) = \mathbf{0}, \tag{4}$$

where $g \in V'$ is some suitable functional satisfying $\langle \phi, g \rangle \neq 0$, and, for $m = 0$, ϕ is sought in $V^0 = V^{O(2)}$. The result is that the original problem can be divided up into a series of problems with reduced complexity. Throughout the rest of this article we shall assume the physically more interesting case where the steady state solutions first become unstable with $m \neq 0$, in which case there is a symmetry breaking bifurcation, rather than just a turning point or fold point.

3 A Posteriori Error Estimation

In this section we consider the derivation of a computable a posteriori estimate for the error in the computed bifurcation point when the extended system (4) is numerically approximated by a general Galerkin finite element method. We begin by

first introducing a suitable finite element approximation of the bifurcation problem (4). To this end, we consider a sequence of $O(2)$ symmetric finite element spaces $V^0_{h,p}$ and finite elements spaces $V^m_{h,p}$ consisting of piecewise polynomial functions of degree p on a partition $\mathcal{T}_h = \{\kappa\}$ of granularity h, from which we shall approximate the $O(2)$ symmetric steady solution and the mth (symmetric or antisymmetric) null-function, respectively.

We find the triple $\mathbf{u}_h = (u_h, \phi_h, \lambda_h) \in \mathbf{V}_{h,p} := V^0_{h,p} \times V^m_{h,p} \times \mathbb{R}$, $m = 1, 2, \ldots$ such that

$$\begin{aligned}\mathcal{N}(\mathbf{u}_h; \mathbf{v}_h) := \hat{\mathcal{N}}(u_h, \lambda_h; \phi_h, v_h) + \hat{\mathcal{N}}'_u(u_h, \lambda_h; \phi_h, \varphi_h) \\ + \chi_h((g, \phi_h) - 1) = 0 \quad \forall \mathbf{v}_h \in \mathbf{V}_{h,p},\end{aligned} \tag{5}$$

where $\mathbf{v}_h := (v_h, \varphi_h, \chi_h)$, $(\cdot, \cdot)$ denotes the standard L^2-inner product, $\hat{\mathcal{N}}(\cdot;\cdot)$ is the semi-linear form associated with the discretization of the underlying steady state version of the partial differential equation (1) and $\hat{\mathcal{N}}'_u(\cdot, \cdot; \cdot, \cdot)$ is the Jacobian of $\hat{\mathcal{N}}(\cdot;\cdot)$ with respect to u and thus represents the discretization of $F'_u(\cdot, \cdot; \cdot)$. For the numerical experiment presented in Sect. 4, this discretization is based on employing the (symmetric) version of the interior penalty discontinuous Galerkin method, together with a Lax–Friedrichs numerical flux approximation of the nonlinear convective terms. In particular, we employ mixed–order spaces of discontinuous piecewise polynomials for the approximation of the velocity and pressure variables present in both the state variable and null-function; for complete details of this numerical scheme, we refer to our recent article [7].

For the proceeding error analysis we make the assumption that (5) is consistent, that is, the analytical solution $\mathbf{u}$ of (4) satisfies

$$\mathcal{N}(\mathbf{u}, \mathbf{v}_h) = 0 \quad \forall \mathbf{v}_h \in \mathbf{V}_{h,p}. \tag{6}$$

For a linear target functional of practical interest $J(\cdot)$, we briefly outline the key steps involved in estimating the approximation error $J(\mathbf{u}) - J(\mathbf{u}_h)$ employing the DWR technique. We write $\mathcal{M}(\cdot, \cdot; \cdot, \cdot)$ to denote the mean value linearization of $\mathcal{N}(\cdot;\cdot)$, defined by

$$\begin{aligned}\mathcal{M}(\mathbf{u}, \mathbf{u}_h; \mathbf{u} - \mathbf{u}_h, \mathbf{w}) &= \mathcal{N}(\mathbf{u}; \mathbf{w}) - \mathcal{N}(\mathbf{u}_h; \mathbf{w}) \\ &= \int_0^1 \mathcal{N}'_{\mathbf{u}}(\theta\mathbf{u} + (1-\theta)\mathbf{u}_h; \mathbf{u} - \mathbf{u}_h, \mathbf{w})\, \mathrm{d}\theta,\end{aligned} \tag{7}$$

for some $\mathbf{w} \in \hat{\mathbf{V}}$. Here, $\hat{\mathbf{V}}$ is some suitably chosen space such that $\mathbf{V}_{h,p} \subset \hat{\mathbf{V}}$. We now introduce the following (formal) *dual problem*: find $\mathbf{z} \in \hat{\mathbf{V}}$ such that

$$\mathcal{M}(\mathbf{u}, \mathbf{u}_h; \mathbf{w}, \mathbf{z}) = J(\mathbf{w}) \quad \forall \mathbf{w} \in \hat{\mathbf{V}}. \tag{8}$$

We assume that (8) possesses a unique solution. This assumption is, of course, dependent on both the definition of $\mathcal{M}(\mathbf{u}, \mathbf{u}_h; \cdot, \cdot)$ and the target functional under

consideration. For the proceeding error analysis, we must therefore assume that (8) is well–posed. By using the linearity of $J(\cdot)$, combining (7) and (8) and using the consistency condition (6) we arrive at the following error representation formula

$$\begin{aligned} J(\mathbf{u}) - J(\mathbf{u}_h) &= J(\mathbf{u} - \mathbf{u}_h) = \mathscr{M}(\mathbf{u}, \mathbf{u}_h; \mathbf{u} - \mathbf{u}_h, \mathbf{z}) \\ &= \mathscr{M}(\mathbf{u}, \mathbf{u}_h; \mathbf{u} - \mathbf{u}_h, \mathbf{z} - \bar{\mathbf{z}}_h) \\ &= -\mathscr{N}(\mathbf{u}_h, \mathbf{z} - \bar{\mathbf{z}}_h) \quad \forall \bar{\mathbf{z}}_h \in \mathbf{V}_{h,p}. \end{aligned} \tag{9}$$

In practice (9) must be numerically estimated by computing a suitable approximation $\mathbf{z}_h$ to the dual solution $\mathbf{z}$. To this end, we compute $\mathbf{z}_h \in \mathbf{V}_{h,\hat{p}}$ using polynomials of degree $\hat{p} > p$ on the same finite element mesh $\mathscr{T}_h$ employed for the primal problem. Furthermore, we point out that the error representation formula (9) may be rewritten in the following equivalent elementwise form

$$J(\mathbf{u}) - J(\mathbf{u}_h) = \sum_{\kappa \in \mathscr{T}_h} \eta_\kappa;$$

here, η_k represent local elemental error indicators, which can be employed within an adaptive refinement strategy, cf. below.

In our case we are interested in controlling the error in the critical bifurcation parameter and hence the target functional of interest is simply $J(\mathbf{u}) = \lambda$. As the dual problem involves the true solution $\mathbf{u}$ we must commit a linearization error and use the approximate $\mathbf{u}_h$ instead. More precisely, we make the following approximation

$$\mathscr{M}(\mathbf{u}, \mathbf{u}_h; \cdot, \cdot) \approx \mathscr{M}(\mathbf{u}_h, \mathbf{u}_h; \cdot, \cdot) = \mathscr{N}'_{\mathbf{u}}(\mathbf{u}_h; \cdot, \cdot).$$

Thereby, the dual problem we actually solve for estimating the error in the approximate critical parameter is given by: find $\mathbf{z}_h := (z_u, z_\phi, z_\lambda) \in \mathbf{V}_{h,\hat{p}}$ such that

$$\begin{aligned} &\hat{\mathscr{N}}'_u(u_h, \lambda_h; v_h, z_u) + \hat{\mathscr{N}}'_\lambda(u_h, \lambda_h; z_u)\chi_h \\ &+ \hat{\mathscr{N}}''_{uu}(u_h, \lambda_h; \varphi_h, \phi_h, z_\phi) + \hat{\mathscr{N}}'_u(u_h, \lambda_h; \varphi_h, z_\phi) \\ &\quad + \hat{\mathscr{N}}''_{u\lambda}(u_h, \lambda_h; \phi_h, z_\phi)\chi_h + z_\lambda(g, \varphi_h) = \chi_h \quad \forall \mathbf{v}_h \in \mathbf{V}_{h,\hat{p}}, \end{aligned} \tag{10}$$

where $\mathbf{v}_h := (v_h, \varphi_h, \chi_h)$.

4 Numerical Experiment

In this section we present a numerical example to demonstrate the practical performance of the proposed a posteriori error estimator derived in Sect. 3 within an automatic *hp*–adaptive refinement procedure which is based on employing one-irregular

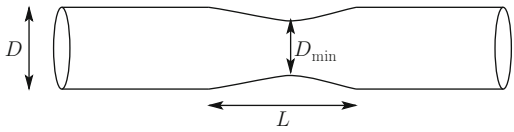


Fig. 1 Stenosis domain

quadrilateral elements. Here, the elements are marked for refinement/derefinement on the basis of the size of the elemental error indicators $|\eta_\kappa|$, using the fixed fraction refinement algorithm with refinement and derefinement fractions set to 25 % and 10 %, respectively. Once an element $\kappa \in \mathcal{T}_h$ has been flagged for refinement, a decision must be made whether the local mesh size or the local degree of the approximating polynomial should be adjusted accordingly. The choice to perform either h– or p–refinement is based on estimating the local smoothness of the (unknown) analytical solution. To this end, we employ the hp–adaptive strategy developed in [12], where the local regularity of the analytical solution is estimated from truncated local Legendre expansions of the computed numerical solution.

We consider a cylindrical pipe of diameter D with an axisymmetric stenotic region of axial length L and radius $r(z)$, given by

$$r(z) = (D_{\min} + (D - D_{\min}) \sin^2(\pi z L))/2, \quad -1/2 \le z/L \le 1/2,$$

where z denotes the coordinate direction along the pipe, centered in the middle of the stenosis, see Fig. 1. Writing S to denote the stenosis degree, defined by $S = 1 - (D_{\min}/D)^2$, we consider the geometry specified by $S = 0.75$, with the stenosis length L/D equal to 2. This problem has been considered recently by Sherwin and Blackburn [13], Blackburn et al. [2] and also as a test problem in Cliffe et al. [4, 7]. In this setting, with a Poiseuille flow profile at the inlet, a steady $O(2)$ symmetry breaking occurs with azimuthal wave number $m = 1$ when $Re^c \approx 721.0527$. We use an initial mesh fitted to the stenosis with 3,840 elements, which is long enough to ensure Poiseuille flow has redeveloped at the outlet of the pipe, and carry out five hp–adaptive refinement steps using the fixed fraction refinement strategy. To this end, Table 1 shows the number of elements, the number of degrees of freedom in computing the primal base solution and the primal null function, the computed critical Reynolds number Re_h^c, the error in the critical Reynolds number, the computed error representation formula $|\sum_{\kappa\in\mathcal{T}_h} \eta_\kappa|$ and the resulting effectivity indices $\tau = |\sum_{\kappa\in\mathcal{T}_h} \eta_\kappa|/|Re^c - Re_h^c|$. We notice immediately that, as the mesh is hp–refined, the effectivity indices tend to unity, indicating that our error indicator is performing extremely well.

In Fig. 2 we plot the results shown in Table 1; in particular, we plot the error $|Re^c - Re_h^c|$ using hp–refinement against the square root of the total number of degrees of freedom employed for both the base and null functions, on a linear-log scale. We see that after the initial pre-asymptotic region, the error in the computed Re_h^c using hp–refinement becomes a straight line, thereby indicating exponential

Table 1 Performance of the hp–adaptive algorithm

No. eles	Base DOF	Null DOF	Re_h^c	$\lvert Re^c - Re_h^c\rvert$	$\lvert\sum_{\kappa\in\mathcal{T}_h}\eta_\kappa\rvert$	τ
3,840	84,480	119,040	662.66203	58.390	67.559	1.16
6,498	146,362	206,090	708.96275	12.090	11.296	0.93
7,512	193,518	271,480	716.36055	4.692	4.680	1.00
9,576	259,439	363,362	721.02371	2.881E − 02	3.575E − 02	1.24
10,101	327,537	456,501	721.05195	5.710E − 04	8.054E − 04	1.41
10,788	398,569	553,522	721.05247	5.477E − 05	5.477E − 05	1.00

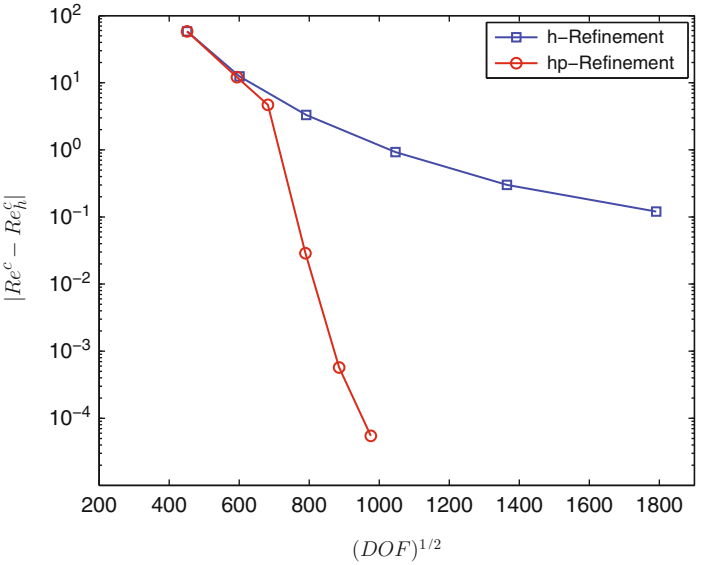


Fig. 2 Comparison between h– and hp–adaptive mesh refinement

convergence. Furthermore, in Fig. 2 we plot the true error in the computed Re_h^c using h–refinement, cf. [7]; here, we clearly observe the superiority of the adaptive hp–refinement algorithm. Indeed, on the final mesh the true error in the computed Re_h^c using hp–refinement is almost four orders of magnitude smaller than the error $|Re^c - Re_h^c|$ when h–refinement is employed alone.

Finally, in Fig. 3 we show the resultant hp–mesh after five adaptive refinement steps. We notice that refinement has been carried out primarily downstream from the stenosis near the wall of the pipe, although some further refinement has also been performed upstream from the stenosis. In particular, we observe that h–refinement has occurred mainly near the boundary of computational domain, as well as in the center of the stenosis, while p–refinement has been utilized elsewhere.

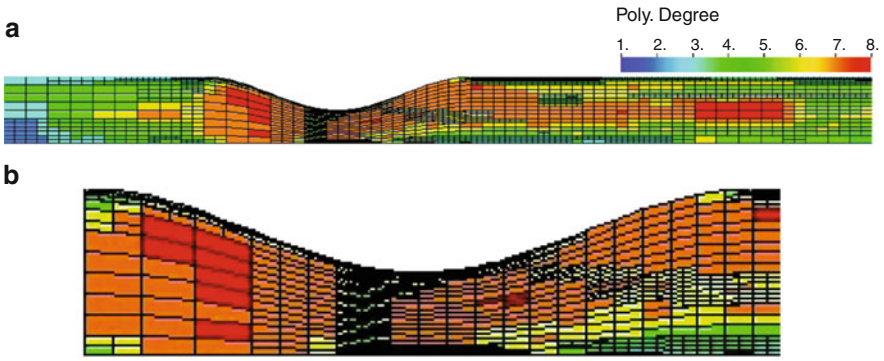


Fig. 3 (**a**) hp–mesh after five adaptive refinements; (**b**) Zoom of (**a**)

References

1. Aston, P.: Analysis and computation of symmetry-breaking bifurcation and scaling laws using group theoretic methods. SIAM J. Math. Anal. **22**, 139–152 (1991)
2. Blackburn, H., Sherwin, S., Barkley, D.: Convective instability and transient growth in steady and pulsatile stenotic flows. J. Fluid Mech. **607**, 267–277 (2008)
3. Brezzi, F., Rappaz, J., Raviart, P.: Finite dimensional approximation of non-linear problems .3. Simple bifurcation points. Numer. Math. **38**(1), 1–30 (1981)
4. Cliffe, K., Hall, E., Houston, P.: Adaptive discontinuous Galerkin methods for eigenvalue problems arising in incompressible fluid flows. SIAM J. Sci. Comput. **31**, 4607–4632 (2010)
5. Cliffe, K., Hall, E., Houston, P., Phipps, E., Salinger, A.: Adaptivity and a posteriori error control for bifurcation problems I: The Bratu problem. Commun. Comput. Phys. **8**, 845–865 (2010)
6. Cliffe, K., Hall, E., Houston, P., Phipps, E., Salinger, A.: Adaptivity and a posteriori error control for bifurcation problems II: Incompressible fluid flow in open systems with Z_2 symmetry. J. Sci. Comput. **47**(3), 389–418 (2011)
7. Cliffe, K., Hall, E., Houston, P., Phipps, E., Salinger, A.: Adaptivity and a posteriori error control for bifurcation problems III: Incompressible fluid flow in open systems with $O(2)$ symmetry. J. of Sci. Comput. **52**(1), 153–179 (2012). In press
8. Cliffe, K., Spence, A., Tavener, S.: O(2)-symmetry breaking bifurcation: with application to the flow past a sphere in a pipe. Internat. J. Numer. Methods Fluids **32**, 175–200 (2000)
9. Cockburn, B., Kanschat, G., Schötzau, D.: The local discontinuous Galerkin method for the Oseen equations. Math. Comp. **73**, 569–593 (2004)
10. Cockburn, B., Kanschat, G., Schötzau, D., Schwab, C.: Local discontinuous Galerkin methods for the Stokes system. SIAM J. Numer. Anal. **40**, 319–343 (2002)
11. Golubitsky, M., Schaeffer, D.: Singularities and Groups in Bifurcation Theory, Vol I. Springer, New York (1985)
12. Houston, P., Süli, E.: A note on the design of hp–adaptive finite element methods for elliptic partial differential equations. Comput. Methods Appl. Mech. Engrg. **194(2–5)**, 229–243 (2005)
13. Sherwin, S., Blackburn, H.: Three–dimensional instabilities and transition of steady and pulsatile axisymmetric stenotics flows. J. Fluid Mech. **533**, 297–327 (2005)
14. Werner, B., Spence, A.: The computation of symmetry-breaking bifurcation points. SIAM J. Numer. Anal. **21**, 388–399 (1984)

hp–Adaptive Two-Grid Discontinuous Galerkin Finite Element Methods for Quasi-Newtonian Fluid Flows

S. Congreve, P. Houston, and T.P. Wihler

Abstract We develop the a posteriori error analysis, with respect to a mesh–dependent energy norm, of two-grid hp–version discontinuous Galerkin finite element methods for quasi-Newtonian flows. The performance of the proposed estimators within an hp–adaptive refinement procedure is studied through a numerical experiment.

1 Introduction

In this paper we develop the a posteriori error analysis, with respect to a mesh-dependent energy norm, of the two-grid hp–version discontinuous Galerkin finite element method (DGFEM) for the quasi-Newtonian fluid flow problem:

$$-\nabla \cdot \left\{ \mu \left(\mathbf{x}, |\mathsf{e}(\mathbf{u})| \right) \mathsf{e}(\mathbf{u}) \right\} + \nabla p = \mathbf{f} \qquad \text{in } \Omega, \tag{1}$$

$$\nabla \cdot \mathbf{u} = 0 \qquad \text{in } \Omega, \tag{2}$$

$$\mathbf{u} = \mathbf{0} \qquad \text{on } \Gamma. \tag{3}$$

Here, $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, is a bounded polygonal or polyhedral Lipschitz domain with boundary $\Gamma = \partial\Omega$, $\mathbf{f} \in L^2(\Omega)^d$, $\mathsf{e}_{ij}(\mathbf{u}) = 1/2 \left(\partial u_i / \partial x_j + \partial u_j / \partial x_i \right)$, $i, j = 1, \ldots, d$, is the symmetric $d \times d$ strain tensor, and $|\mathsf{e}(\mathbf{u})|$ is the Frobenius norm of $\mathsf{e}(\mathbf{u})$. We assume that $\mu \in \mathrm{C}(\bar{\Omega} \times [0, \infty))$ and there exists constants $m_\mu, M_\mu > 0$ such that

S. Congreve (✉) · P. Houston
School of Mathematical Sciences, University of Nottingham, Nottingham, NG7 2RD, UK
e-mail: pmxsc@nottingham.ac.uk; Paul.Houston@nottingham.ac.uk

T.P. Wihler
Mathematisches Institut, Universität Bern, Sidlerstrasse 5, CH-3012, Bern, Switzerland
e-mail: wihler@math.unibe.ch

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_37,

$$m_\mu(t-s) \le \mu(\mathbf{x},t)t - \mu(\mathbf{x},s)s \le M_\mu(t-s)\,, \qquad t \ge s \ge 0\,, \quad \mathbf{x} \in \bar{\Omega}. \tag{4}$$

For ease of notation we write $\mu(t)$ instead of $\mu(\mathbf{x},t)$.

Two-grid methods were first introduced by Xu [8–10] for the continuous Galerkin FEM; for related work, see [5], and the references cited therein. Recent developments [2, 5] have focused on the analysis of the two-grid method for the DGFEM for second-order quasilinear partial differential equations. Two-grid methods only require a non-linear system of equations to be solved on a coarse mesh and a linearised system on the fine mesh; thereby, potentially leading to significant computational savings. In this paper we extend the a posteriori error analysis presented in [3,5] to the two-grid DGFEM for (1)–(3). Here, DGFEM is considered due to the ease in which it can handle hp-adaptive mesh refinements. Throughout this paper, we use the following standard function spaces. For a bounded Lipschitz domain $D \subset \mathbb{R}^d, d \ge 1$, we write $H^t(D)$ to denote the usual Sobolev space of real-valued functions of order $t \ge 0$ with norm $\|\cdot\|_{t,D}$. In the case when $t = 0$, we set $L^2(D) = H^0(D)$. We define $H_0^1(D)$ to be the subspace of functions in $H^1(D)$ with zero trace on ∂D. Additionally, we set $L_0^2(D) = \{q \in L^2(D) : \int_D q\,\mathrm{d}\mathbf{x} = 0\}$.

The outline of this article is as follows. Section 2 introduces the two-grid DGFEM for the numerical approximation of (1)–(3). In Sect. 3 we state the a posteriori error bound for the scheme and in Sect. 4 we outline the hp–adaptive mesh refinement strategy. Finally, in Sect. 5 we validate the results with a numerical experiment.

2 Two-Grid hp–Version DGFEM

We consider shape-regular meshes $\mathscr{T}_h$ that partition Ω into open disjoint parallelograms (in $\mathbb{R}^2$) or hexahedra (in $\mathbb{R}^3$) κ such that $\overline{\Omega} = \bigcup_{\kappa\in\mathscr{T}_h} \overline{\kappa}$. By h_κ we denote the element diameter of $\kappa \in \mathscr{T}_h$ and let $\mathbf{n}_\kappa$ signify the unit outward normal vector to κ. Further, we define $\mathbf{h} = \{h_\kappa : \kappa \in \mathscr{T}_h\}$. We allow the meshes $\mathscr{T}_h$ to be '1-irregular'; moreover, we suppose that $\mathscr{T}_h$ is 'regularly reducible' [7]. Note that these assumptions imply the family $\{\mathscr{T}_h\}_{h>0}$ is of 'bounded local variation', i.e., there exists a constant $\rho_1 \ge 1$, independent of element sizes, such that $\rho_1^{-1} \le h_\kappa/h_{\kappa'} \le \rho_1$ for any pair of elements $\kappa, \kappa' \in \mathscr{T}_h$ which share a common face $F = \partial\kappa \cap \partial\kappa'$.

For a non-negative integer k, we denote by $\mathscr{Q}_k(\kappa)$ the set of all tensor-product polynomials of degree k in each variable on κ. To each $\kappa \in \mathscr{T}_h$, we assign a polynomial degree $k_\kappa \ge 1$ and store these in a vector $\mathbf{k} = \{k_\kappa : \kappa \in \mathscr{T}_h\}$. We suppose that $\mathbf{k}$ is also of bounded local variation, i.e., there exists a constant $\rho_2 \ge 1$, independent of the element sizes and $\mathbf{k}$, such that, for any pair of neighbouring elements $\kappa, \kappa' \in \mathscr{T}_h$, $\rho_2^{-1} \le k_\kappa/k_{\kappa'} \le \rho_2$. With this notation we introduce the finite element spaces

$$\mathbf{V}(\mathscr{T}_h,\mathbf{k}) = \left\{\mathbf{v}\in L^2(\Omega)^d : \mathbf{v}|_\kappa \in \mathscr{Q}_{k_\kappa}(\kappa)^d, \kappa\in\mathscr{T}_h\right\},$$

$$Q(\mathscr{T}_h,\mathbf{k}) = \left\{q\in L^2_0(\Omega) : q|_\kappa \in \mathscr{Q}_{k_\kappa-1}(\kappa), \kappa\in\mathscr{T}_h\right\}.$$

We define an interior face ((d−1) dimensional facet) F of $\mathscr{T}_h$ as the intersection of two neighbouring elements $\kappa,\kappa'\in\mathscr{T}_h$, i.e., $F=\partial\kappa\cap\partial\kappa'$. Similarly, we define a boundary face $F\subset\Gamma$ as the face of an element κ on the boundary. We denote by $\mathscr{F}_h^{\mathscr{I}}$ the set of all interior faces, $\mathscr{F}_h^{\mathscr{B}}$ the set of all boundary faces and $\mathscr{F}_h = \mathscr{F}_h^{\mathscr{I}}\cup\mathscr{F}_h^{\mathscr{B}}$ the set of all faces.

We shall now define some suitable face operators that are required for the definition of the proceeding DGFEM. Let q, $\mathbf{v}$, and $\boldsymbol{\tau}$ be scalar-, vector- and matrix-valued functions, respectively, which are smooth inside each element $\kappa\in\mathscr{T}_h$. Given two adjacent elements, $\kappa^+,\kappa^-\in\mathscr{T}_h$, which share a common face $F\in\mathscr{F}_h^{\mathscr{I}}$, i.e., $F=\partial\kappa^+\cap\partial\kappa^-$, we write $q^\pm$, $\mathbf{v}^\pm$, and $\boldsymbol{\tau}^\pm$ to denote the traces of the functions q, $\mathbf{v}$, and $\boldsymbol{\tau}$, respectively, on the face F, taken from the interior of $\kappa^\pm$, respectively. The averages of q, $\mathbf{v}$ and $\boldsymbol{\tau}$ at $\mathbf{x}\in F$ are given by $\{\!\{q\}\!\} = {}^1\!/_2(q^+ + q^-)$, $\{\!\{\mathbf{v}\}\!\} = {}^1\!/_2(\mathbf{v}^+ + \mathbf{v}^-)$, and $\{\!\{\boldsymbol{\tau}\}\!\} = {}^1\!/_2(\boldsymbol{\tau}^+ + \boldsymbol{\tau}^-)$, respectively. Similarly, the jumps of q, $\mathbf{v}$ and $\boldsymbol{\tau}$ at $\mathbf{x}\in F$ are given by $[\![q]\!] = q^+\mathbf{n}_{\kappa^+} + q^-\mathbf{n}_{\kappa^-}$, $[\![\mathbf{v}]\!] = \mathbf{v}^+\cdot\mathbf{n}_{\kappa^+} + \mathbf{v}^-\cdot\mathbf{n}_{\kappa^-}$, $\underline{[\![\mathbf{v}]\!]} = \mathbf{v}^+\otimes\mathbf{n}_{\kappa^+} + \mathbf{v}^-\otimes\mathbf{n}_{\kappa^-}$, and $[\![\boldsymbol{\tau}]\!] = \boldsymbol{\tau}^+\mathbf{n}_{\kappa^+} + \boldsymbol{\tau}^-\mathbf{n}_{\kappa^-}$. On a boundary face $F\in\mathscr{F}_h^{\mathscr{B}}$, we set $\{\!\{q\}\!\} = q$, $\{\!\{\mathbf{v}\}\!\} = \mathbf{v}$, $\{\!\{\boldsymbol{\tau}\}\!\} = \boldsymbol{\tau}$, $[\![q]\!] = q\mathbf{n}$, $[\![\mathbf{v}]\!] = \mathbf{v}\cdot\mathbf{n}$, $\underline{[\![\mathbf{v}]\!]} = \mathbf{v}\otimes\mathbf{n}$ and $[\![\boldsymbol{\tau}]\!] = \boldsymbol{\tau}\mathbf{n}$, with $\mathbf{n}$ denoting the unit outward normal vector on Γ. For a face $F\in\mathscr{F}_h$, we define h_F as the face diameter and the polynomial degree k_F as $k_F=\max(k_\kappa,k_{\kappa'})$, if $F=\partial\kappa\cap\partial\kappa'\in\mathscr{F}_h^{\mathscr{I}}$, and $k_F=k_\kappa$, if $F=\partial\kappa\cap\Gamma\in\mathscr{F}_h^{\mathscr{B}}$.

With this notation, we first introduce the so-called *standard* (interior penalty) DGFEM for the numerical approximation of the problem (1)–(3). To this end, given a (fine) mesh $\mathscr{T}_h$ partition of Ω, together with a corresponding polynomial degree vector $\mathbf{k}$, the DGFEM is defined as follows: find $(\mathbf{u}_{h,k},p_{h,k})\in\mathbf{V}(\mathscr{T}_h,\mathbf{k})\times Q(\mathscr{T}_h,\mathbf{k})$ such that

$$A_{h,k}(\mathbf{u}_{h,k};\mathbf{u}_{h,k},\mathbf{v}_{h,k}) + B_{h,k}(\mathbf{v}_{h,k},p_{h,k}) = F_{h,k}(\mathbf{v}_{h,k}), \tag{5}$$

$$-B_{h,k}(\mathbf{u}_{h,k},q_{h,k}) = 0 \tag{6}$$

for all $(\mathbf{v}_{h,k},q_{h,k})\in\mathbf{V}(\mathscr{T}_h,\mathbf{k})\times Q(\mathscr{T}_h,\mathbf{k})$, where

$$\begin{aligned}A_{h,k}(\varphi;\mathbf{u},\mathbf{v}) =& \int_\Omega \mu(|\mathsf{e}_h(\varphi)|)\mathsf{e}_h(\mathbf{u}):\mathsf{e}_h(\mathbf{v})\,\mathrm{d}\mathbf{x} - \sum_{F\in\mathscr{F}_h}\int_F \{\!\{\mu(|\mathsf{e}_h(\varphi)|)\mathsf{e}_h(\mathbf{u})\}\!\} : \underline{[\![\mathbf{v}]\!]}\,\mathrm{d}s\\ &- \sum_{F\in\mathscr{F}_h}\int_F \{\!\{\mu(h_F^{-1}|\underline{[\![\boldsymbol{\varphi}]\!]}|)\mathsf{e}_h(\mathbf{v})\}\!\} : \underline{[\![\mathbf{u}]\!]}\,\mathrm{d}s + \sum_{F\in\mathscr{F}_h}\int_F \sigma_{h,k}\underline{[\![\mathbf{u}]\!]} : \underline{[\![\mathbf{v}]\!]}\,\mathrm{d}s,\end{aligned}$$

$$B_{h,k}(\mathbf{v},q) = -\int_\Omega q\nabla_h\cdot\mathbf{v}\,\mathrm{d}\mathbf{x} + \sum_{F\in\mathscr{F}_h}\int_F \{\!\{q\}\!\}[\![\mathbf{v}]\!]\,\mathrm{d}s, \qquad F_{h,k}(\mathbf{v}) = \int_\Omega \mathbf{f}\cdot\mathbf{v}\,\mathrm{d}\mathbf{x}.$$

Here, $\mathsf{e}_h(\cdot)$ and ∇_h denote the broken strain tensor and gradient operator, respectively, defined elementwise. For a face $F \in \mathscr{F}_h$ the *interior penalty parameter* $\sigma_{h,k}$ is defined as $\sigma_{h,k} = \gamma k_F^2 h_F^{-1}$, where $\gamma > 0$ is a constant. We note that, due to the condition on the non-linearity (4), the interior penalty stabilisation may be selected independent of $\mu(\cdot)$, provided the penalty parameter is chosen sufficiently large (independent of local element sizes and polynomial degrees) such that it is greater than a constant $\gamma_{\min}$ dependent on m_μ and M_μ; see [3]. The non-linearity present in the face terms of the DGFEM can be handled in an easy fashion, since the structure of the underlying linearised system of equations stemming from the application of a Newton-type non-linear iterative solver is identical to that arising in the linear case.

Given the above scheme, we may now proceed to introduce the two-grid DGFEM approximation to (1)–(3). To this end, we consider a fine and coarse partition $\mathscr{T}_h$ and $\mathscr{T}_H$, respectively, of the computational domain Ω. Here, we define $\mathbf{H} = \{h_\kappa : \kappa \in \mathscr{T}_H\}$. In particular, we assume that $\mathscr{T}_h$ and $\mathscr{T}_H$ are nested in the sense that, for any $\kappa_h \in \mathscr{T}_h$ there exists an element $\kappa_H \in \mathscr{T}_H$ such that $\bar{\kappa}_h \subseteq \bar{\kappa}_H$. Moreover, to each mesh $\mathscr{T}_h$ and $\mathscr{T}_H$, we associate a corresponding polynomial degree distribution $\mathbf{k} = \{k_\kappa : \kappa \in \mathscr{T}_h\}$ and $\mathbf{K} = \{K_\kappa : \kappa \in \mathscr{T}_H\}$, respectively, with the property that, given $\kappa_h \in \mathscr{T}_h$ and the associated $\kappa_H \in \mathscr{T}_H$, such that $\bar{\kappa}_h \subseteq \bar{\kappa}_H$, the corresponding polynomial degrees satisfy the condition $k_{\kappa_h} \geq K_{\kappa_H}$. With this notation, we now introduce the hp–version of the two-grid algorithm [2] for the DGFEM discretisation of (1)–(3):

1. (Non-linear solve) Compute $(u_{H,K}, p_{H,K}) \in \mathbf{V}(\mathscr{T}_H, \mathbf{K}) \times Q(\mathscr{T}_H, \mathbf{K})$ such that

$$A_{H,K}(\mathbf{u}_{H,K}; \mathbf{u}_{H,K}, \mathbf{v}_{H,K}) + B_{H,K}(\mathbf{v}_{H,K}, p_{H,K}) = F_{H,K}(\mathbf{v}_{H,K}) \,, \tag{7}$$

$$-B_{H,K}(\mathbf{u}_{H,K}, q_{H,K}) = 0 \tag{8}$$

   for all $(v_{H,K}, q_{H,K}) \in \mathbf{V}(\mathscr{T}_H, \mathbf{K}) \times Q(\mathscr{T}_H, \mathbf{K})$.
2. (Linear solve) Find the fine grid solution $(u_{2G}, p_{2G}) \in \mathbf{V}(\mathscr{T}_h, \mathbf{k}) \times Q(\mathscr{T}_h, \mathbf{k})$, i.e.,

$$A_{h,k}(\mathbf{u}_{H,K}; \mathbf{u}_{2G}, \mathbf{v}_{h,k}) + B_{h,k}(\mathbf{v}_{h,k}, p_{2G}) = F_{h,k}(\mathbf{v}_{h,k}) \,, \tag{9}$$

$$-B_{h,k}(\mathbf{u}_{2G}, q_{h,k}) = 0 \tag{10}$$

   for all $(v_{h,k}, q_{h,k}) \in \mathbf{V}(\mathscr{T}_h, \mathbf{k}) \times Q(\mathscr{T}_h, \mathbf{k})$.

We note that due to the construction of the fine and coarse meshes we have that $\mathbf{V}(\mathscr{T}_H, \mathbf{K}) \subseteq \mathbf{V}(\mathscr{T}_h, \mathbf{k})$ and $Q(\mathscr{T}_H, \mathbf{K}) \subseteq Q(\mathscr{T}_h, \mathbf{k})$.

3 A Posteriori Error Estimates

In this section, we recall a posteriori error bounds for the two-grid DGFEM defined by (7)–(10). To this end, we introduce the following energy norms

$$\|\mathbf{v}\|_{h,k}^2 = \|\mathsf{e}_h(\mathbf{v})\|_{0,\Omega}^2 + \sum_{F\in\mathscr{F}_h}\int_F \sigma_{h,k}|\underline{[\![\mathbf{v}]\!]}|^2\,\mathrm{d}s\,, \qquad \|(\mathbf{v},q)\|_{DG}^2 = \|\mathbf{v}\|_{h,k}^2 + \|q\|_{0,\Omega}^2\,,$$

on the fine mesh $\mathscr{T}_h$ with polynomial degree vector $\mathbf{k}$. Writing Π_κ to denote the L^2–projection operator onto $\mathbf{V}(\mathscr{T}_h,\mathbf{k})$, we now state the following upper bound.

Theorem 1. *Let* $(\mathbf{u},p)\in H_0^1(\Omega)^d\times L_0^2(\Omega)$ *be the analytical solution of* (1)–(3), $(\mathbf{u}_{H,K},p_{H,K})\in\mathbf{V}(\mathscr{T}_H,\mathbf{K})\times Q(\mathscr{T}_H,\mathbf{K})$ *the numerical approximation obtained from* (7)–(8) *and* $(\mathbf{u}_{2G},p_{2G})\in\mathbf{V}(\mathscr{T}_h,\mathbf{k})\times Q(\mathscr{T}_h,\mathbf{k})$ *the numerical approximation computed from* (9)–(10)*; then the following hp–a posteriori error bound holds*

$$\|(\mathbf{u}-\mathbf{u}_{2G},p-p_{2G})\|_{DG} \le C\left(\sum_{\kappa\in\mathscr{T}_h}\left(\eta_\kappa^2+\xi_\kappa^2\right)+\sum_{\kappa\in\mathscr{T}_h}h_\kappa^2k_\kappa^{-2}\,\|\mathbf{f}-\Pi_\kappa\mathbf{f}\|_{0,\kappa}^2\right)^{\frac{1}{2}}, \tag{11}$$

with a constant $C>0$*, which is independent of* $\mathbf{h}$, $\mathbf{H}$, $\mathbf{k}$ *and* $\mathbf{K}$*. Here, for all* $\kappa\in\mathscr{T}_h$*, the* local fine grid error indicators η_κ *are defined by*

$$\begin{aligned}\eta_\kappa^2 &= h_\kappa^2k_\kappa^{-2}\left\|\Pi_{\kappa,k_\kappa}\mathbf{f}+\nabla\cdot\{\mu(|\mathsf{e}(\mathbf{u}_{H,K})|)\mathsf{e}(\mathbf{u}_{2G})\}\right\|_{0,\kappa}^2+\|\nabla\cdot\mathbf{u}\|_{0,\kappa}^2 \\ &\quad+h_\kappa k_\kappa^{-1}\left\|[\![p_{2G}]\!]-[\![\mu(|\mathsf{e}_h(\mathbf{u}_{H,K})|)\mathsf{e}_h(\mathbf{u}_{2G})]\!]\right\|_{0,\partial\kappa\setminus\Gamma}^2+\gamma^2h_\kappa^{-1}k_\kappa^3\left\|\underline{[\![\mathbf{u}_{2G}]\!]}\right\|_{0,\partial\kappa^2}\end{aligned} \tag{12}$$

and the local two-grid error indicators ξ_κ *are defined, for all* $\kappa\in\mathscr{T}_h$*, as*

$$\xi_\kappa^2 = \left\|\left(\mu\left(|\mathsf{e}(\mathbf{u}_{H,K})|\right)-\mu\left(|\mathsf{e}(\mathbf{u}_{2G})|\right)\right)\mathsf{e}(\mathbf{u}_{2G})\right\|_{0,\kappa}^2. \tag{13}$$

Proof. The proof is based on exploiting the techniques developed in [3, 5]; for full details, we refer to [4].

4 Two-Grid hp–Adaptive Mesh Refinement Algorithm

For the standard DGFEM discretisation of the non-Newtonian problem (1)–(3), the mesh may be automatically constructed using the hp–adaptive refinement algorithm outlined in [3]. In that setting, the local error indicators are defined in an analogous way to η_κ given in (12), with $\mathbf{u}_{H,P}$ and $\mathbf{u}_{2G}$ both replaced by $\mathbf{u}_{h,p}$. In the context of the two-grid DGFEM discretisation defined by (7)–(10), it is necessary to refine both the fine and coarse meshes, together with their corresponding polynomial degrees, in order to decrease the error measured in the energy norm.

In [5] we proposed an algorithm that refined the fine mesh based only on η_κ and the coarse mesh based only on ξ_κ. In this article we propose an alternative refinement strategy based on first employing $\eta_\kappa+\xi_\kappa$ to identify regions where refinement needs

to be performed; the relative size of η_κ and ξ_κ is then used to indicate which of the two meshes should be refined. This algorithm is outlined below.

Algorithm 4.1. *The finite element spaces* $\mathbf{V}(\mathscr{T}_h, \mathbf{k}) \times Q(\mathscr{T}_h, \mathbf{k})$ *and* $\mathbf{V}(\mathscr{T}_H, \mathbf{K}) \times Q(\mathscr{T}_H, \mathbf{K})$ *are constructed, based on employing the following algorithm.*

0. *Initial step: Select initial coarse and fine meshes* $\mathscr{T}_H$ *and* $\mathscr{T}_h$*, as well as initial coarse and fine polynomial degree distributions* $\mathbf{K}$ *and* $\mathbf{k}$*, respectively, in such a manner that* $\mathbf{V}(\mathscr{T}_H, \mathbf{K}) \subseteq \mathbf{V}(\mathscr{T}_h, \mathbf{k})$ *and* $Q(\mathscr{T}_H, \mathbf{K}) \subseteq Q(\mathscr{T}_h, \mathbf{k})$.
1. *Select elements of* $\mathscr{T}_h$ *and* $\mathscr{T}_H$ *for refinement/derefinement, as follows:*
   (a) *Determine the sets* $R(\mathscr{T}_h) \subseteq \mathscr{T}_h$ *and* $D(\mathscr{T}_h) \subseteq \mathscr{T}_h$ *of fine elements to be (potentially) refined/derefined, respectively, based on the size of* $\eta_\kappa + \xi_\kappa$ *using a standard refinement algorithm, e.g., the fixed fraction refinement strategy.*
   (b) *Derefine all fine elements* $\kappa \in D(\mathscr{T}_h)$.
   (c) *For all elements selected for refinement decide whether to perform refinement of the fine or coarse mesh: for all* $\kappa \in R(\mathscr{T}_h)$
      - *If* $\lambda \xi_\kappa \leq \eta_\kappa$ *refine the fine element* κ*, and*
      - *If* $\lambda \eta_\kappa \leq \xi_\kappa$ *refine the coarse element* $\kappa_H \in \mathscr{T}_H$*, where* $\kappa \subseteq \kappa_H$.

      Here, $\lambda \in (0, 1]$ *is a steering parameter. We note that both of these statements can be true, in which case both coarse and fine elements are refined.*
2. *For elements marked for refinement in the fine and coarse mesh, determine whether to perform h– or p–refinement; see, for example, [6].*
3. *Smooth the mesh to ensure that* $\mathbf{V}(\mathscr{T}_H, \mathbf{K}) \subseteq \mathbf{V}(\mathscr{T}_h, \mathbf{k})$ *and* $Q(\mathscr{T}_H, \mathbf{K}) \subseteq Q(\mathscr{T}_h, \mathbf{k})$.

Remark 1. For the purposes of this article we initially set $\mathbf{V}(\mathscr{T}_H, \mathbf{K}) = \mathbf{V}(\mathscr{T}_h, \mathbf{k})$ and $Q(\mathscr{T}_H, \mathbf{K}) = Q(\mathscr{T}_h, \mathbf{k})$ in Step 0 above.

5 Numerical Experiment

In this section we present a numerical experiment to validate the performance of the a posteriori error bound derived in Theorem 1 within the automatic two-grid hp–adaptive mesh refinement procedure defined in Algorithm 4.1 based on 1-irregular quadrilateral elements for $\Omega \subset \mathbb{R}^2$. In step 1 of Algorithm 4.1 we employ the fixed fraction strategy with refinement and derefinement fractions set to 25 % and 5 %, respectively. We note here that we start with a polynomial degree of $k_K = 3$ for all elements in both the coarse and fine mesh. We set the interior penalty parameter constant $\gamma = 10$ and the steering parameter $\lambda = 0.5$. As well as exploiting the two-grid method, we compute the standard DGFEM formulation (5) and (6) for comparison.

We consider the cavity-like problem from [1, Sect. 6.1] using the Carreau law non-linearity $\mu(|\mathsf{e}(\mathbf{u})|) = 1 + (1 + 5|\mathsf{e}(\mathbf{u})|^2)^{-0.1}$. We let $\Omega = (0, 1)^2 \subset \mathbb{R}^2$ be the

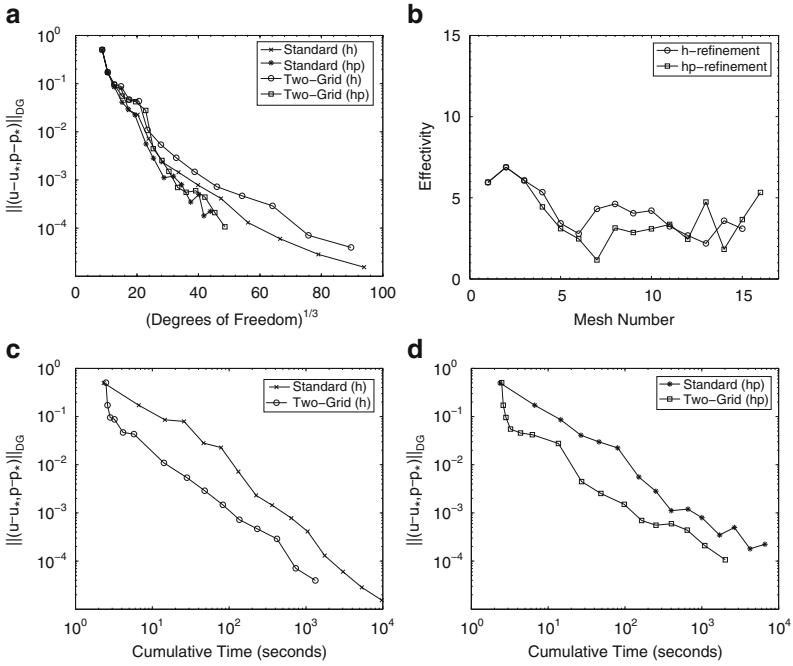


Fig. 1 (**a**) Comparison of the error in the DG norm, using the standard non-linear solver ($\mathbf{u}_* = \mathbf{u}_{h,p}, p_* = p_{h,p}$) and the two-grid method ($\mathbf{u}_* = \mathbf{u}_{2G}, p_* = p_{2G}$), with respect to the number of degrees of freedom; (**b**) Effectivity of the h– and hp–refinement using the two-grid method; (**c**) and (**d**) Comparison of the error in the DG norm, using both the standard non-linear solver ($\mathbf{u}_* = \mathbf{u}_{h,p}, p_* = p_{h,p}$) and the two-grid method ($\mathbf{u}_* = \mathbf{u}_{2G}, p_* = p_{2G}$), with respect to cpu time for the h– and hp– refinement strategies, respectively

unit square and select the forcing function $\mathbf{f}$ so that

$$\mathbf{u}(x,y) = \begin{pmatrix} \left(1-\cos\left(2\pi\,\frac{e^{1.8x}-1}{e^{1.8}-1}\right)\right)\sin(2\pi y) \\ -1.8e^{1.8x}\sin\left(2\pi\,\frac{e^{1.8x}-1}{e^{1.8}-1}\right)\frac{1-\cos(2\pi y)}{e^{1.8}-1} \end{pmatrix},$$

$$p(x,y) = 3.6\pi e^{1.8x}\sin\left(2\pi\,\frac{e^{1.8x}-1}{e^{1.8}-1}\right)\frac{\sin(2\pi y)}{e^{1.8}-1}.$$

In Fig. 1a we present a comparison of the actual error measured in terms of the energy norm versus the third root of the number of degrees of freedom (of the fine mesh) for both the standard DGFEM formulation (5) and (6), together with the two-grid DGFEM (7)–(10). In this figure we perform both h– and hp–adaptive mesh refinement for both schemes. Here, we can see that, for the problem at hand, the true error in the two-grid DGFEM is only marginally worse than the corresponding quantity for the standard DGFEM, when the same number of degrees of freedom in the two-grid fine mesh, as in the mesh for the standard DGFEM, are used. From

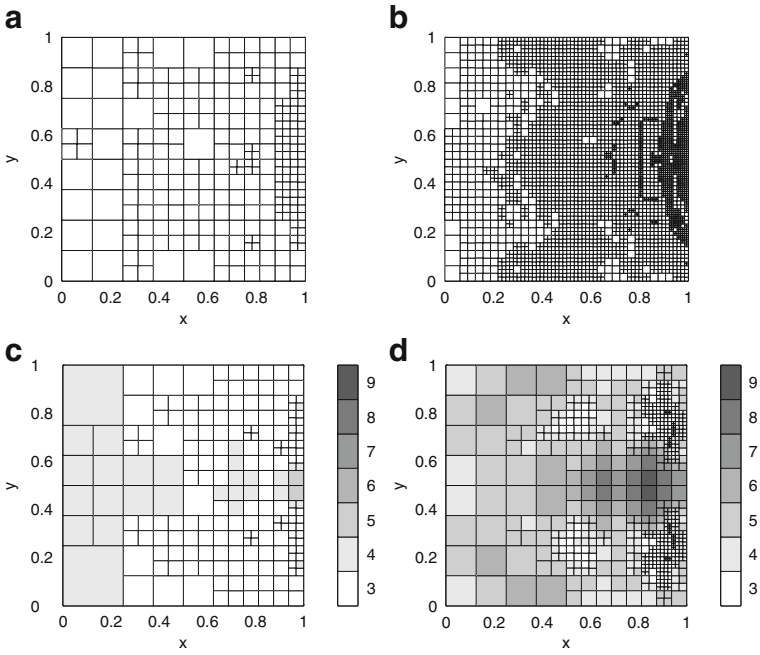


Fig. 2 (**a**) and (**b**) The coarse mesh and fine mesh, respectively, after 11 h–adaptive refinement steps; (**c**) and (**d**) The coarse mesh and fine mesh, respectively, after 11 hp–adaptive refinement steps

Fig. 1b, we observe that for both the h– and hp–refinement strategy, the error bound overestimates the true error by roughly a consistent amount, in the sense that the effectivity indices are roughly constant; indeed, here, the effectivity indices are around 4. We note that although the two-grid DGFEM gives a slightly worse error than the standard DGFEM, for a fixed number of fine mesh degrees of freedom, the latter method is computationally more expensive. Indeed, the error, measured in the DGFEM norm, for both the standard and two-grid methods, when both h– and hp–adaptive mesh refinement has been employed, compared to the cumulative cpu time required for the calculation of each numerical solution is shown in Figs. 1c, d. These figures clearly illustrate the superiority of employing the two-grid DGFEM.

In Fig. 2 we show the fine and coarse h– and hp–refinement meshes after 11 mesh refinements. For h–refinement, we can see that all the fine grid refinement occurs mostly around the 'hill' and 'valley' in the pressure on the right of the domain, as would occur for the standard DGFEM. Notice that only a small amount of refinement has taken place in the corresponding elements in the coarse mesh, namely, wherever ξ_κ is expected to be large. In the fine mesh generated employing hp–refinement, h–refinement occurs mostly around the center of the hills and valleys in the pressure, with p–refinement being utilised elsewhere.

Acknowledgements PH acknowledges the financial support of the EPSRC under the grant EP/H005498. TW acknowledges the financial support of the SNF No. 200021_126594.

References

1. S. Berrone and E. Süli. Two-sided a posteriori error bounds for incompressible quasi-Newtonian flows. *IMA J. Numer. Anal.*, 28:382–421, 2008.
2. C. Bi and V. Ginting. Two-grid discontinuous Galerkin method for quasi-linear elliptic problems. *J. Sci. Comput.*, 49(3):311–331, 2011.
3. S. Congreve, P. Houston, E. Süli, and T. P. Wihler. Discontinuous Galerkin finite element approximation of quasilinear elliptic boundary value problems II: Strongly monotone quasi-Newtonian flows.
4. S. Congreve, P. Houston, E. Süli, and T. P. Wihler. Two-grid hp-version discontinuous Galerkin finite element methods for quasi-Newtonian flows. (In preparation).
5. S. Congreve, P. Houston, and T. P. Wihler. Two-grid hp-version discontinuous Galerkin finite element methods for second-order quasilinear elliptic PDEs. J. Sci. Comput. (in press)
6. P. Houston and E. Süli. A note on the design of hp-adaptive finite element methods for elliptic partial differential equations. *Comput. Methods Appl. Mech. Engrg.*, 194(2–5):229–243, 2005.
7. C. Ortner and E Süli. Discontinuous Galerkin finite element approximation of nonlinear second-order elliptic and hyperbolic systems. *SIAM J. Numer. Anal.*, 45(4):1370–1397, 2007.
8. J. Xu. A new class of iterative methods for nonselfadjoint or indefinite problems. *SIAM J. Numer. Anal.*, 29(2):303–319, 1992.
9. J. Xu. A novel two-grid method for semilinear elliptic equations. *SIAM J. Sci. Comput.*, 15(1):231–237, 1994.
10. J. Xu. Two-grid discretization techniques for linear and nonlinear PDEs. *SIAM J. Numer. Anal.*, 33(5):1759–1777, 1996.

Discontinuous Galerkin Methods for Eigenvalue Problems on Anisotropic Meshes

E.J.C. Hall and S. Giani

Abstract We derive a goal-oriented a posteriori error estimate for hp-adaptive discontinuous Galerkin discretizations of convection-diffusion eigenvalue problems. We consider one-irregular meshes consisting of parallelograms. The estimate yields very accurate measurements of the errors in the two target functionals considered in this paper. The accuracy of our error estimator is also confirmed by the effectivity index very close to 1 in all numerical tests. We apply our goal-oriented estimator as an error indicator in an anisotropic hp-adaptive refinement algorithm and illustrate its practical performance in a series of numerical examples.

1 Introduction

In this paper, we derive a goal-oriented a posteriori error estimator for a discontinuous Galerkin (DG) discretization of convection-diffusion eigenvalue problems. It is well-known that eigenfunctions to such problems may have boundary or internal layers of small width where their gradients change extremely rapidly. One way to deal efficiently with boundary or internal layers is to use adaptive finite element methods that are capable of locally refining the meshes and adjusting the order of the elements in the vicinity of these layers. DG methods are particularly appropriate for adaptivity because they provide increased flexibility in mesh design (irregular grids are admissible) and the freedom to choose the elemental polynomial degrees, without the need to enforce continuity between elements. For further information regarding DG methods see, for example, [1, 9–11].

E.J.C. Hall · S. Giani
School of Mathematical Sciences, University of Nottingham, University Park, Nottingham, NG7 2RD, UK
e-mail: edward.hall@nottingham.ac.uk; stefano.giani@nottingham.ac.uk

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_38,

There is now quite a substantial literature on a posteriori error estimation for eigenvalue problems; in the context of energy norm error estimators we refer the reader to [4, 6, 7, 12], while goal oriented a posteriori error estimates have been developed in [2,3,8]. One of the main advantages of the goal oriented approach is the possibility to choose which measurement of the eigenpairs to target with adaptivity. In order to illustrate such flexibility, we consider in this paper two different target quantities of the computed eigenpair, namely the eigenvalue and the mean of the eigenfunction. We then formulate an anisotropic adaptive strategy based on the work in [5].

2 Model Problem and DG Discretisation

Let Ω be a bounded open polygonal domain in $\mathbb{R}^2$ and let Γ signify the union of its 1–dimensional open faces. We consider the advection–diffusion–reaction eigenvalue problem:

$$\mathscr{L}u \equiv -\nabla\cdot(a\nabla u) + \nabla\cdot(\mathbf{b}u) + cu = \lambda u \quad \text{in } \Omega\,, \tag{1}$$

$$u = 0 \quad \text{on } \Gamma, \tag{2}$$

where $c \in L^\infty(\Omega)$ is real–valued, $\mathbf{b} = \{b_i\}_{i=2}^d$ is a vector function whose entries b_i are Lipschitz continuous real–valued functions on $\bar{\Omega}$, and $a = \{a_{ij}\}_{i,j=1}^2$ is a *symmetric* matrix whose entries a_{ij} are bounded, piecewise continuous real–valued functions defined on $\bar{\Omega}$, with

$$\boldsymbol{\zeta}^\top a(x)\boldsymbol{\zeta} \geq 0 \quad \forall \boldsymbol{\zeta} \in \mathbb{R}^2\,, \quad \text{a.e. } x \in \bar{\Omega}\,. \tag{3}$$

The standard weak formulation of (1) and (2) is to find the eigenpair $(\lambda, u) \in \mathbb{C} \times H_0^1(\Omega)$ such that

$$A(u,\bar{v}) \equiv \int_\Omega (a\nabla u\cdot\nabla\bar{v} + \nabla(\mathbf{b}\cdot u)\bar{v} + cu\bar{v})\,dx = \lambda(u,\bar{v}) \quad \forall v \in H_0^1(\Omega), \tag{4}$$

where the space $H_0^1(\Omega)$ is the standard space of (complex) functions with gradient in $L^2(\Omega)$ with zero trace on Γ and $(\cdot,\cdot)$ is the standard $L^2(\Omega)$ inner product. By normalising the eigenfunction so that $(u,\bar{u}) = 1$, the eigenvalue problem can be rewritten as find $\hat{u} := (\lambda, u) \in V := \mathbb{C} \times H_0^1(\Omega)$ such that

$$\mathscr{N}(\hat{u},\hat{v}) = 0, \quad \forall \hat{v} := (\chi, v) \in V, \tag{5}$$

where

$$\mathscr{N}(\hat{u},\hat{v}) := \lambda(u,\bar{v}) - A(u,\bar{v}) + \chi((u,\bar{u}) - 1).$$

Let $\mathscr{T}_h = \{\kappa\}$ be a subdivision of the (polygonal) domain Ω into disjoint open element domains κ. Throughout we insist that each κ is constructed via an affine mapping F_κ of the reference element $\hat{\kappa} := (-1,1)^2$ and let $h_{1,\kappa}$ be the length of the image of $(-1,1)\times\{-1\}$ under the mapping F_κ and similarly let $h_{2,\kappa}$ be the length of the image of $\{-1\}\times(-1,1)$ under F_κ. The elements are assumed to be anisotropic in the sense that there is no condition imposed between $h_{1,\kappa}$ and $h_{2,\kappa}$.

The model problem (1) and (2) is discretized using the symmetric interior penalty DG (SIPG) method presented in [5]. We assume a piecewise polynomial approximation and define the DG finite element space $S_{h,p}$ as follows

$$S_{h,p} = \{u \in L^2(\Omega) \,:\, u|_\kappa \circ F_\kappa \in \mathscr{Q}_p(\hat{\kappa});\ \kappa \in \mathscr{T}_h\},$$

where $\mathscr{Q}_p(\hat{\kappa})$ is the set of complex multiples of tensor-product polynomials on $\hat{\kappa}$ of degree p in each coordinate direction.

Let our approximate eigenfunctions, u_{DG} be normalised so that $(u_{\mathrm{DG}}, \bar{u}_{\mathrm{DG}}) = 1$, then the SIPG method can be written in the form: find $\hat{u}_{\mathrm{DG}} := (\lambda_h, u_{\mathrm{DG}}) \in \mathbb{C}\times S_{h,p} := V_{h,p}$ such that

$$\mathscr{N}_{\mathrm{DG}}(\hat{u}_{\mathrm{DG}}; \hat{v}) = 0, \quad \forall \hat{v} := (\chi, \hat{v}) \in V_{h,p}, \tag{6}$$

where

$$\mathscr{N}_{\mathrm{DG}}(\hat{u}_{\mathrm{DG}}; \hat{v}) := \lambda_h(u_{\mathrm{DG}}, \bar{v}) - A_{\mathrm{DG}}(u_{\mathrm{DG}}, \bar{v}) + \chi((u_{\mathrm{DG}}, \bar{u}_{\mathrm{DG}}) - 1), \tag{7}$$

and $A_{\mathrm{DG}}(\cdot,\cdot)$ is the SIPG bilinear form.

3 A Posteriori Error Estimation

In this section we consider the derivation of an a posteriori error estimate for a given eigenvalue of interest. To this end, we first recall the DWR technique in the context of error estimation for general target functionals of the solution; for further details, see [2].

3.1 DWR Approach for Functionals

For a general linear target functional of practical interest $J(\cdot) : V \to \mathbb{R}$, we briefly outline the key steps involved in estimating the approximation error $J(\hat{u}) - J(\hat{u}_{\mathrm{DG}})$ employing the DWR technique. We introduce the mean value linearization of $\mathscr{N}_{\mathrm{DG}}(\cdot;\cdot)$ by

$$\mathscr{M}(\hat{u}, \hat{u}_{\mathrm{DG}}; \hat{u} - \hat{u}_{\mathrm{DG}}, \hat{w}) = \mathscr{N}_{\mathrm{DG}}(\hat{u}, \hat{w}) - \mathscr{N}_{\mathrm{DG}}(\hat{u}_{\mathrm{DG}}, \hat{w})$$
$$= \int_0^1 \mathscr{N}_{\hat{u}}'[\theta\hat{u} + (1-\theta)\hat{u}_{\mathrm{DG}}](\hat{u} - \hat{u}_{\mathrm{DG}}, \hat{w})\,\mathrm{d}\theta, \qquad (8)$$

where $\mathscr{N}_{\hat{u}}'[\hat{w}](\cdot, \hat{v})$ denotes the Fréchet derivative of $\mathscr{N}(\cdot, \hat{v})$ evaluated at some $\hat{w} \in V$ We now introduce the following (formal) *dual problem*: find $\hat{z} := (\lambda_z, z) \in V$ such that

$$\mathscr{M}(\hat{u}, \hat{u}_{\mathrm{DG}}; \hat{v}, \hat{z}) = J(\hat{v}) \quad \forall \hat{v} \in V. \qquad (9)$$

We assume that (9) possesses a unique solution. This assumption is, of course, dependent on both the definition of $\mathscr{M}(\hat{u}, \hat{u}_{\mathrm{DG}}; \cdot, \cdot)$ and the target functional under consideration. For the proceeding error analysis, we must therefore assume that (9) is well–posed.

Proposition 1 (Error Representation Formula). *Let $\hat{u}$ and $\hat{u}_{\mathrm{DG}}$ denote the solutions of (5) and (6) respectively and suppose that the dual problem (9) is well posed with solution $\hat{z}$. Then*

$$J(\hat{u}) - J(\hat{u}_{\mathrm{DG}}) = -\mathscr{N}_{\mathrm{DG}}(\hat{u}_{\mathrm{DG}}, \hat{z} - \hat{z}_h) = \sum_{\kappa \in \mathscr{T}_h} \eta_\kappa \quad \forall \hat{z}_h \in V_{h,p}.$$

See [3] for a proof.

In general we must approximate the dual solution $\hat{z}$, but we cannot approximate from $V_{h,p}$ due to (6), instead we uniformly increase the polynomial degree by 1 and find $\hat{z}_{\mathrm{DG}} := (\lambda_z, z_{\mathrm{DG}}) \in V_{h,p+1}$. We also linearize about our approximate solution $\hat{u}_{\mathrm{DG}}$ and hence our dual problem becomes: find $\hat{z}_{\mathrm{DG}} \in V_{h,p+1}$ such that

$$\lambda_h(v, \bar{z}_{\mathrm{DG}}) - A_{\mathrm{DG}}(v, \bar{z}_{\mathrm{DG}}) + \chi(u_{\mathrm{DG}}, \bar{z}_{\mathrm{DG}}) + 2\bar{\lambda}_z \mathrm{Re}(v, \bar{u}_{\mathrm{DG}}) = J(\hat{v}) \quad \forall \hat{v} \in V_{h,p+1}. \qquad (10)$$

We consider two different functionals:

1. $J(\hat{u}) = \lambda$ and so enables us to evaluate the error committed in our approximation of the eigenvalue. In this case the approximate dual problem becomes: find $\hat{z}_{\mathrm{DG}} \in V_{h,p+1}$ such that

$$\lambda_h(v, \bar{z}_{\mathrm{DG}}) - A_{\mathrm{DG}}(v, \bar{z}_{\mathrm{DG}}) + \chi(u_{\mathrm{DG}}, \bar{z}_{\mathrm{DG}}) + 2\bar{\lambda}_z \mathrm{Re}(v, \bar{u}_{\mathrm{DG}}) = \chi \quad \forall \hat{v} \in V_{h,p+1}. \qquad (11)$$

   We notice here that this formulation forces the dual solution to be scaled so that $(u_{\mathrm{DG}}, \bar{z}_{\mathrm{DG}}) = 1$.
2. $J(\hat{u}) = (u, w)$, where $w \in L^2(\Omega)$ is some weighting function. Now the approximate dual problem becomes: find $\hat{z}_{\mathrm{DG}} \in V_{h,p+1}$ such that

$$\lambda_h(v, \bar{z}_{\mathrm{DG}}) - A_{\mathrm{DG}}(v, \bar{z}_{\mathrm{DG}}) + \chi(u_{\mathrm{DG}}, \bar{z}_{\mathrm{DG}}) + 2\bar{\lambda}_z \mathrm{Re}(v, \bar{u}_{\mathrm{DG}}) = (v, w) \quad \forall \hat{v} \in V_{h,p+1}. \qquad (12)$$

   Here we notice that the dual solution is scaled so that $(u_{\mathrm{DG}}, \bar{z}_{\mathrm{DG}}) = 0$.

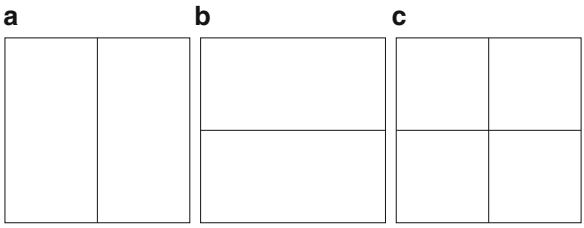


Fig. 1 Cartesian refinement in 2D: (**a**) and (**b**) Anisotropic refinement; (**c**) Isotropic refinement

4 Anisotropic Adaptive Strategy

Suppose we are interested in finding $\hat{u}_{\mathrm{DG}}$ such that $|J(\hat{u}) - J(\hat{u}_{\mathrm{DG}})| < \mathrm{Tol}$, where Tol is some tolerance. We enforce this in practice by ensuring the stopping criterion

$$|\mathscr{N}_{\mathrm{DG}}(\hat{u}_{\mathrm{DG}}, \hat{z}_{\mathrm{DG}} - \hat{z}_h)| \leq \mathrm{Tol}\,. \tag{13}$$

If (13) is not satisfied, then elements are marked for refinement/derefinement according to the size of the (approximate) error indicators $|\eta_\kappa|$ in (10). In our numerical experiments we use a fixed fraction mesh refinement algorithm, with refinement and derefinement fractions set to 20 % and 10 %, respectively.

To subdivide the elements which have been flagged for refinement, we employ a simple Cartesian refinement strategy; here, elements may be subdivided either anisotropically or isotropically according to the three refinements depicted in Fig. 1. In order to determine the optimal refinement, we propose the following strategy based on choosing the most competitive subdivision of κ from a series of trial refinements, whereby an approximate local error indicator on each trial patch is determined.

Given an element κ, on each mesh patch, $\mathscr{T}_{h,i}$, $i = 1, 2, 3$ from Fig. 1 we compute the approximate error estimators

$$\mathscr{N}_{\kappa,i}(u_{\mathrm{DG},i}, \hat{z}_i - z_h) = \sum_{\kappa' \in \mathscr{T}_{h,i}} \eta_{\kappa',i}\,,$$

for $i = 1, 2, 3$, respectively. Here, $u_{\mathrm{DG},i}$, $i = 1, 2, 3$, is the discontinuous Galerkin approximation to (1) and (2) computed on the mesh patch $\mathscr{T}_{h,i}$, $i = 1, 2, 3$, respectively. Similarly $\hat{z}_i$ denotes the discontinuous Galerkin approximation to $\hat{z}$. Details of how this is carried out can be found in [5].

The element κ is then refined according to the subdivision of κ which satisfies

$$\min_{i=1,2,3} \frac{|\eta_\kappa| - |\mathscr{N}_{\kappa,i}(u_{\mathrm{DG},i}, \hat{z}_i - \hat{z}_h)|}{\#\mathrm{dofs}(\mathscr{T}_{h,i}) - \#\mathrm{dofs}(\kappa)},$$

where $\#\mathrm{dofs}(\kappa)$ and $\#\mathrm{dofs}(\mathscr{T}_{h,i})$, $i = 1, 2, 3$, denote the number of degrees of freedom associated with κ and $\mathscr{T}_{h,i}$, $i = 1, 2, 3$, respectively.

Remark 1. It is not possible to solve the local version of the (primal) eigenvalue problem as an eigenvalue problem because we have to enforce nonzero boundary conditions. Instead we set the eigenvalue on our local patch to be our approximate eigenvalue λ_h and solve a linear system by removing the now inconsistent scaling $(u_{\mathrm{DG}}, \bar{u}_{\mathrm{DG}}) = 1$. Similarly, for the local dual problem, the (functional dependent) scaling of z_{DG} is inconsistent with the imposition of local boundary conditions. Hence we fix λ_z and solve a linear system only for z_i.

5 Numerical Experiment

In our numerical experiment we choose $\Omega = (0,1)^2$, pick $a = \epsilon I$, $\mathbf{b} = (-1,-1)^\top$, $c = 1$. In this situation we know that all (true) eigenvalues are real valued and are given by $\lambda_{m,n} = \epsilon\pi^2(m^2+n^2)+(1+2\epsilon)/(2\epsilon)$, while the (unscaled) eigenfunctions are given by

$$u_{m,n} = \mathrm{e}^{-(x+y)/2\epsilon} \sin(m\pi x)\sin(n\pi y).$$

1. $J(\hat{u}) = \lambda$. We select $\epsilon = 0.05$ and let $m, n = 1$. In Fig. 2 we plot the true eigenvalue solution and corresponding dual solution, noticing that the primal eigenfunction has boundary layers in the bottom lefthand corner, while boundary layers appear in the top righthand corner in the dual solution. This eigenfunction seems well chosen to highlight the effectivity of our anisotropic strategy. We select $p = 1$ and choose an initial mesh of 64 square elements and compare our anisotropic adaptive strategy against a more standard isotropic strategy. Table 1 shows the errors and effectivities for the anisotropic strategy, while Fig. 3a shows a plot of errors against degrees of freedom for both the anisotropic and isotropic strategies, while the mesh after six anisotropic refinement steps is given in Fig. 3b.

   Firstly we notice that the error estimator is performing very well, with effectivities close to unity on the final meshes. However, we see that, except on the final mesh, the anisotropic refinement strategy is only a little better than the isotropic strategy. Interestingly, if we look at the refined mesh, the majority of refinement has been carried out in the center of the domain where there is little structure in either the primal or dual solutions and isotropic refinement has been selected. This seems counter intuitive, but the a posteriori estimation technique weights the primal residual with the dual solution and even though the residual may be large in the bottom left hand corner, the dual solution is negligible there. It is interesting that in order to obtain a good approximation to an eigenvalue the corresponding eigenfunction need barely be resolved at all. We note that some refinement has been performed near the boundary and here the anisotropy of the solutions has been picked out.

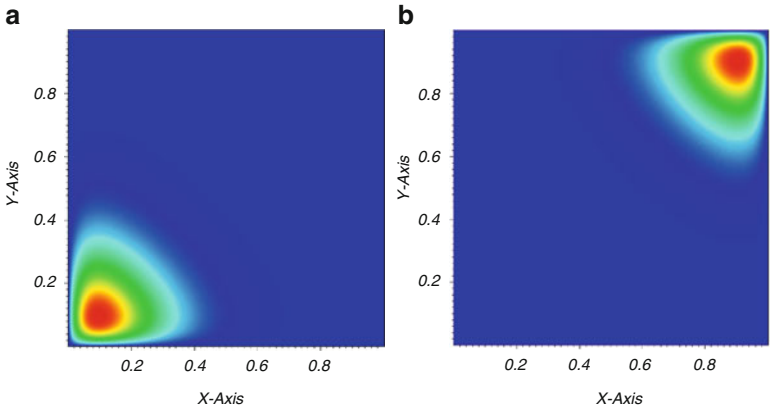


Fig. 2 (**a**) Primal eigenfunction (**b**) dual solution

Table 1 Error convergence and effectivities, $J(\hat{u}) = \lambda$

Mesh	No. eles.	No. DOF	λ	$\lvert\lambda - \lambda_h\rvert$	Effectivity
1	64	256	12.96921	0.9822	1.65
2	84	336	12.91810	0.9311	1.22
3	129	516	12.31893	0.3320	2.14
4	188	752	12.07141	8.4454E − 02	1.29
5	309	1,236	12.03095	4.3994E − 02	1.13
6	489	1,956	12.02161	3.4654E − 02	1.07
7	800	3,200	12.00797	2.1005E − 02	1.00

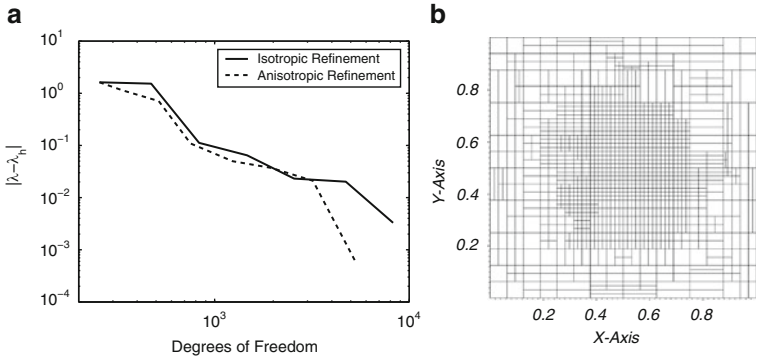


Fig. 3 (**a**) Error convergence and (**b**) mesh after six anisotropic mesh refinement steps, $J(\hat{u}) = \lambda$

Table 2 Error convergence and effectivities, $J(\hat{u}) = (u, w)$

Mesh	No. eles.	No. DOF	λ	$\lvert J(\hat{u}) - J(\hat{u}_{\mathrm{DG}})\rvert$	Effectivity
1	64	256	12.96921	5.0168E − 02	0.33
2	85	340	12.78572	3.9947E − 02	0.91
3	117	468	12.26553	2.1198E − 02	1.57
4	160	640	12.06717	9.6751E − 03	1.18
5	239	956	12.09901	5.9610E − 03	1.17
6	349	1,396	12.04475	3.5773E − 03	1.07
7	531	2,124	12.01486	2.0940E − 03	1.03
8	804	3,216	12.01322	1.5415E − 03	1.00

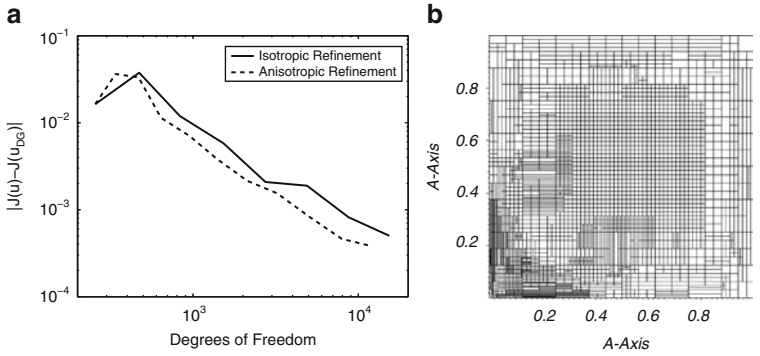


Fig. 4 (**a**) Error convergence and (**b**) mesh after seven anisotropic mesh refinement steps, $J(\hat{u}) = (u, w)$

2. $J(\hat{u}) = (u, 1)$. Suppose $\epsilon = 0.05$, then $(u, 1) = 0.364100985230600$. Again we compare our anisotropic strategy against an isotropic strategy beginning with a mesh consisting of 64 square elements and $p = 1$. The dual solution in this situation is very similar in appearance to the dual solution for the eigenvalue error, but has a different scaling.

   Table 2 shows the errors and effectivities for the anisotropic strategy, while Fig. 4a shows an error convergence plot for both the anisotropic and isotropic strategies, while the mesh after seven anisotropic refinement steps is given in Fig. 4b.

   Once again we see effectivities close to 1 on the final meshes, indicating our error indicator is performing very well. For this functional we do see an improvement over isotropic refinement when using our anisotropic strategy, albeit a small one, but one which is consistent on all meshes. We notice now that much more refinement has been carried out in the bottom left hand corner where there is structure in the primal solution and the anisotropic nature of the refinement is as expected. Again isotropic refinement has been performed in the center of the domain, slightly against intuition.

References

1. Arnold, D. N., Brezzi, F., Cockburn, B., Marini, L. D.: Unified analysis of discontinuous Galerkin methods for elliptic problems. SIAM J. Numer. Anal. **39**, 1749–1779, (2001/2).
2. Becker, R. and Rannacher, R.: An optimal control approach to a-posteriori error estimation in finite element methods. In: Iserles, A. (eds.) Acta Numerica, pp. 1–102. CUP, (2001).
3. Cliffe, K.A., Hall, E.J.C. and Houston, P.: Adaptive Discontinuous Galerkin Methods for Eigenvalue Problems arising in Incompressible Fluid Flows. SIAM J. Sci. Comput. **31**, 4607–4632 (2010).
4. R.G. Durán, R.G., Padra, C. and Rodriguez, R.:A posteriori error estimates for the finite element approximation of eigenvalue problems. Math. Models Methods Appl. Sci. **13**:1219–1229, (2003).
5. Georgoulis, E.H., Hall E.J.C. and Houston, P.: Discontinuous Galerkin Methods for Advection-Diffusion-Reaction Problems on Anisotropically Refined Meshes. SIAM J. Sci. Comput. **30**, 246–271 (2007).
6. Giani, S. and Graham, I.: A convergent adaptive method for elliptic eigenvalue problems. SIAM J. Numer. Anal. **47**. 1067–1091, (2009).
7. Hall, E.J.C. and Giani, S.: An a posteriori error estimator for hp–adaptive discontinuous Galerkin methods for elliptic eigenvalue problems. Math. Models Methods Appl. Sci., *in press*.
8. Heuveline, V. and Rannacher, R.: A Posteriori Error Control for Finite Element Approximations of Elliptic Eigenvalue Problems. Adv. Comp. Math. **15**, 107–138 (2001).
9. Houston, P., Schwab, Ch. and Süli, E.: Discontinuous hp-finite element methods for advection-diffusion-reaction problems. SIAM J. Numer. Anal. **39**, 2133–2163, (2002).
10. Johnson, C. and Pitkäranta, J.:An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation. Math. Comp. **46**, 1–26, (1986).
11. Reed, W. H. and Hill, T. R.: Triangular mesh methods for the neutron transport equation. Los Alamos Scientific Laboratory Tech. Report LA-UR-73-479 (1973).
12. Walsh, T.F., Reese, G.M. and Hetmaniuk, U.L.: Explicit a posteriori error estimates for eigenvalue analysis of heterogeneous elastic structures. Comput. Methods Appl. Mech. Engrg. **196**, 3614–3623, (2007).

Two Dimensional Compressible Fluid-Structure Interaction Model Using DGFEM

J. Hasnedlová-Prokopová, M. Feistauer, A. Kosík, and V. Kučera

Abstract The subject of this paper is the numerical solution of the interaction of compressible flow and an elastic body with a special emphasis on the simulation of vibrations of vocal folds during phonation onset. The time-dependence of the domain occupied by the fluid is treated by the ALE (Arbitrary Lagrangian-Eulerian) method and the compressible Navier-Stokes equations are written in the ALE form. The deformation of the elastic body, caused by the aeroelastic forces, is described by the linear dynamical elasticity equations. Both these systems are coupled by transmission conditions. For the space-discretization of the flow problem the discontinuous Galerkin finite element method (DGFEM) is used. The time-discretization is realized by the backward difference formula (BDF). The structural problem is discretized by the conforming finite element method and the Newmark method. The results of the use of two different couplings and their comparison are presented.

1 Flow Problem

We are concerned with the problem of compressible flow in a time-dependent bounded domain $\Omega_t \subset \mathbb{R}^2$ with $t \in [0, T]$. 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} denotes impermeable time-dependent walls.

The time-dependence of the domain Ω_t is treated with the use of the *arbitrary Lagrangian-Eulerian* (ALE) method, see [6]. This method is based on a regular one-to-one ALE mapping $\mathcal{A}_t$ of the reference configuration $\overline{\Omega}_0$ onto the current

J. Hasnedlová-Prokopová (✉) · M. Feistauer · A. Kosík · V. Kučera
Department of Numerical Mathematics, Faculty of Mathematics and Physics, Charles University, Sokolovská 83, 18675 Praha 8, Czech Republic
e-mail: jarkaprokop@post.cz; feist@karlin.mff.cuni.cz; adam.kosik@atlas.cz; kucera@karlin.mff.cuni.cz

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_39,

configuration $\overline{\Omega}_t$. In this framework we can write the system of governing equations consisting of the continuity equation, the Navier-Stokes equations and the energy equation in the ALE form

$$\frac{D^{\mathcal{A}}\mathbf{w}}{Dt} + \sum_{s=1}^{2} \frac{\partial \mathbf{g}_s(\mathbf{w})}{\partial x_s} + \mathbf{w}\mathrm{div}\mathbf{z} = \sum_{s=1}^{2} \frac{\partial \mathbf{R}_s(\mathbf{w}, \nabla \mathbf{w})}{\partial x_s}. \tag{1}$$

For a detailed description see e.g. [3]. Here $D^{\mathcal{A}}/Dt$ is the ALE derivative,

$$\begin{aligned}
\mathbf{w} &= (\rho, \rho v_1, \rho v_2, E)^T \in I\!R^4, \qquad (2)\\
\mathbf{g}_s(\mathbf{w}) &= \mathbf{f}_s(\mathbf{w}) - z_s\mathbf{w},\ s = 1,2,\\
\mathbf{f}_s(\mathbf{w}) &= (\rho v_s, \rho v_1 v_s + \delta_{1s}p, \rho v_2 v_s + \delta_{2s}p, (E+p)v_s)^T,\ s = 1,2,\\
\mathbf{R}_s(\mathbf{w}, \nabla \mathbf{w}) &= (0, \tau_{s1}^V, \tau_{s2}^V, \tau_{s1}^V v_1 + \tau_{s2}^V v_2 + k\frac{\partial \theta}{\partial x_s})^T,\ s = 1,2,\\
\tau_{ij}^V &= \lambda \delta_{ij}\mathrm{div}\mathbf{v} + 2\mu d_{ij}(\mathbf{v}),\ d_{ij}(\mathbf{v}) = \frac{1}{2}\left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i}\right),\ i,j = 1,2.
\end{aligned}$$

We use the following notation: $\mathbf{z}$ – domain velocity, ρ – fluid density, p – pressure, E – total energy, $\mathbf{v} = (v_1, v_2)$ – velocity vector, θ – absolute temperature, $c_v > 0$ – specific heat at constant volume, $\gamma > 1$ – Poisson adiabatic constant, $\mu > 0, \lambda = -2\mu/3$ – viscosity coefficients, $k > 0$ – heat conduction coefficient, τ_{ij}^V – components of the viscous part of the stress tensor. The vector-valued functions $\mathbf{w}$ is called state vector, $\mathbf{f}_s$ are inviscid fluxes and $\mathbf{R}_s$ represent viscous terms. The system (1) is completed by the thermodynamical relations

$$p = (\gamma - 1)\left(E - \rho\frac{|\mathbf{v}|^2}{2}\right), \quad \theta = \frac{1}{c_v}\left(\frac{E}{\rho} - \frac{1}{2}|\mathbf{v}|^2\right) \tag{3}$$

and equipped with the initial condition $\mathbf{w}(\mathbf{x}, 0) = \mathbf{w}^0(\mathbf{x}),\ \mathbf{x} \in \Omega_0$ and the boundary conditions

$$\begin{aligned}
&\text{Inlet } \Gamma_I: \quad \rho = \rho_D, \quad \mathbf{v} = \mathbf{v}_D = (v_{D1}, v_{D2}), \qquad (4)\\
&\qquad \sum_{j=1}^{2}\left(\sum_{i=1}^{2}\tau_{ij}^V n_i\right)v_j + k\frac{\partial\theta}{\partial \mathbf{n}} = 0;\\
&\text{Moving wall } \Gamma_{W_t}: \quad \mathbf{v} = \mathbf{z}_D(t) = \text{velocity of a moving wall}, \quad \frac{\partial\theta}{\partial \mathbf{n}} = 0;\\
&\text{Outlet } \Gamma_O: \quad \sum_{j=1}^{2}\tau_{ij}^V n_j = 0, \quad \frac{\partial\theta}{\partial \mathbf{n}} = 0, \quad i = 1,2,
\end{aligned}$$

with prescribed data $\rho_D,\ \mathbf{v}_D,\ \mathbf{z}_D$.

The space discretization of the problem is carried out by the discontinuous Galerkin finite element method. For the time discretization we use a semi-implicit scheme, where the ALE-derivative is approximated by the first-order backward difference and the remaining terms are treated with the aid of a linearization and time extrapolation (see, e.g. [2,4]).

2 Elasticity Problem

By $\Omega^b \subset I\!R^2$ we shall denote a bounded open set representing an elastic body. Elastic deformations of this body are described by the dynamical equations

$$\varrho^b \frac{\partial^2 u_i}{\partial t^2} + C\varrho^b \frac{\partial u_i}{\partial t} - \sum_{j=1}^{2} \frac{\partial \tau_{ij}^b}{\partial x_j} = 0, \quad \text{in } \Omega^b \times (0,T), \quad i = 1,2, \tag{5}$$

where $\mathbf{u}(\mathbf{x},t) = (u_1(\mathbf{x},t), u_2(\mathbf{x},t))$, $\mathbf{x} \in \Omega^b$, $t \in (0,T)$, is the displacement, $(\tau_{ij}^b)_{i,j=1}^2$ represents the stress tensor fulfilling the generalized Hooke law for isotropic material

$$\tau_{ij}^b = \tilde{\lambda} \mathrm{div}\, \mathbf{u} \delta_{ij} + 2\tilde{\mu} e_{ij}, \quad i,j = 1,2, \tag{6}$$

with the Lamé coefficients $\tilde{\lambda}, \tilde{\mu}$. Often the Young modulus E^b and the Poisson ratio σ^b are used. They are defined by $E^b = \frac{\tilde{\mu}(3\tilde{\lambda}+2\tilde{\mu})}{\tilde{\lambda}+\tilde{\mu}}$, $\sigma^b = \frac{\tilde{\lambda}}{2(\tilde{\lambda}+\tilde{\mu})}$. Further, $e_{ij}(\mathbf{u}) = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right)$, $i,j = 1,2$, are the components of the strain tensor and ϱ^b is the density of the solid material. The dissipation of the energy of the system is represented by the expression $C\varrho^b \frac{\partial u_i}{\partial t}$, where $C \geq 0$.

The formulation of the dynamical elasticity problem (5) is completed by the initial conditions $\mathbf{u}(\mathbf{x},0) = 0$ and $\frac{\partial \mathbf{u}}{\partial t}(\mathbf{x},0) = 0$, $\mathbf{x} \in \Omega^b$, and boundary conditions on the boundary $\partial\Omega^b = \Gamma_W^b \cup \Gamma_D^b$, where $\Gamma_W^b \cap \Gamma_D^b = \emptyset$, $\Gamma_W^b \subset \Gamma_{W_0}$ and Γ_D^b is a fixed part of the boundary:

$$\sum_{j=1}^{2} \tau_{ij}^b n_j = T_i^{\mathbf{n}} \quad \text{on } \Gamma_W^b \times (0,T), \quad i = 1,2, \tag{7}$$

$$\mathbf{u} = 0 \quad \text{on } \Gamma_D^b \times (0,T). \tag{8}$$

By $T_i^{\mathbf{n}}$, $i = 1,2$, we denote the components of the normal stress. The space discretization is treated by the finite element method using continuous piecewise linear elements, which leads to a second order system of ordinary differential equations. For the time discretization we use the Newmark scheme. For more detail see e.g. [5].

3 Fluid-Structure Interaction Coupling and the Construction of the ALE Mapping

Up to now the fluid flow and the deformation of the elastic body have been considered as two separate problems. Now we need to take into account the mutual interaction of the fluid and the body on the common boundary

$$\tilde{\Gamma}_{W_t} = \left\{\mathbf{x} \in IR^2;\ \mathbf{x} = \mathbf{X} + \mathbf{u}(\mathbf{X}, t),\ \mathbf{X} \in \Gamma_W^b\right\}.$$

The domain Ω_t is determined by the displacement $\mathbf{u}$ on Γ_W^b at time t. If the domain Ω_t occupied by the fluid at time t is known, we can solve the problem describing the flow and compute the surface force acting on the body on $\tilde{\Gamma}_{W_t}$, which can be transformed to the reference configuration, i.e. to the interface Γ_W^b. In the case of the linear elasticity model, when only small deformations are considered, we get the transmission condition

$$\sum_{j=1}^{2} \tau_{ij}^b(\mathbf{X}) n_j(\mathbf{X}) = -\sum_{j=1}^{2} \tau_{ij}^f(\mathbf{x}) n_j(\mathbf{X}), \quad i = 1, 2, \tag{9}$$

where τ_{ij}^f are the components of the stress tensor of the fluid: $\tau_{ij}^f = -p\delta_{ij} + \tau_{ij}^V$, $i, j = 1, 2$. The points $\mathbf{x}$ and $\mathbf{X}$ satisfy the relation

$$\mathbf{x} = \mathbf{X} + \mathbf{u}(\mathbf{X}, t) \tag{10}$$

and $\mathbf{n}(\mathbf{X}) = (n_1(\mathbf{X}), n_2(\mathbf{X}))$ denotes the unit outer normal to the body Ω^b on Γ_W^b at the point $\mathbf{X}$. Further, the fluid velocity is defined on the moving part of the boundary Γ_{W_t} by the transmission condition

$$\mathbf{v}(\mathbf{x}, t) = \mathbf{z}_D(\mathbf{x}, t) = \frac{\partial \mathbf{u}(\mathbf{X}, t)}{\partial t}. \tag{11}$$

The ALE mapping $\mathscr{A}_t$ is determined with the aid of an artificial stationary elasticity problem, where we seek $\mathbf{d} = (d_1, d_2)$ defined in Ω_0 as a solution of the elastic system

$$\sum_{j=1}^{2} \frac{\partial \tau_{ij}^a}{\partial x_j} = 0 \text{ in } \Omega_0, \quad i = 1, 2, \tag{12}$$

where τ_{ij}^a are the components of the artificial stress tensor $\tau_{ij}^a = \lambda^a \mathrm{div}\mathbf{d}\delta_{i,j} + 2\mu^a e_{ij}^a(\mathbf{d})$, $e_{ij}^a(\mathbf{d}) = \frac{1}{2}\left(\frac{\partial d_i}{\partial x_j} + \frac{\partial d_j}{\partial x_i}\right)$, $i = 1, 2$. The Lamé coefficients λ^a and μ^a are related to the artificial Young modulus E^a and the artificial Poisson number σ_a the same as in Sect. 2. The boundary conditions for $\mathbf{d}$ are prescribed by

$$\mathbf{d}|_{\Gamma_I \cup \Gamma_O} = 0,\ \mathbf{d}|_{\Gamma_{W_0} \setminus \Gamma_W^b} = 0,\ \mathbf{d}(\mathbf{x}, t) = \mathbf{u}(\mathbf{x}, t),\ \mathbf{x} \in \Gamma_W^b. \tag{13}$$

The solution of the problem (12) and (13) gives us the ALE mapping of $\overline{\Omega}_0$ onto $\overline{\Omega}_t$ in the form

$$\mathcal{A}_t(\mathbf{x}) = \mathbf{x} + \mathbf{d}(\mathbf{x}, t), \quad \mathbf{x} \in \overline{\Omega}_0, \tag{14}$$

for each time t.

The numerical solution of the problem (12) and (13) is carried out by the conforming piecewise linear finite elements.

4 Coupling Procedure

In the solution of the complete coupled fluid-structure interaction problem it is necessary to apply a suitable coupling procedure. The general framework can be found, e.g. in [1]. In our case we apply two different types of algorithms. First, the so-called strong coupling will be presented.

1. Assume that the approximate solution of the flow problem and the deformation of the structure $\mathbf{u}_k$ on the time level t_k are known.
2. Set $\mathbf{u}_{k+1}^0 := \mathbf{u}_k$, $l := 1$ and apply the iterative process:
   (a) Compute the stress tensor τ_{ij}^f and the aerodynamical force acting on the structure and transform it to the interface Γ_W^b.
   (b) Solve the elasticity problem, compute the deformation $\mathbf{u}_{k+1}^l$ and the approximation $\Omega_{t_{k+1}}^l$ of the domain occupied by the fluid at time t_{k+1}.
   (c) Determine the ALE mapping $\mathcal{A}_{t_{k+1}}^l$ and approximate the domain velocity $\mathbf{z}_{k+1}^l$.
   (d) Solve the flow problem on the approximation $\Omega_{t_{k+1}}^l$.
   (e) If the variation $\mathbf{u}_{k+1}^l - \mathbf{u}_{k+1}^{l-1}$ of the displacement is larger than the prescribed tolerance, go to (a) and $l := l + 1$. Else $k := k + 1$ and go to (2).

In order to obtain the second type of the algorithm, the weak (loose) coupling, in step (e) we set $k := k + 1$ and go to (2) already in the case when $l = 1$.

5 Numerical Results

The motivation for our numerical experiments is the simulation of vocal folds vibrations during phonation onset.

First, let us present a simplified model of vocal folds consisting of a channel with two bumps (see Fig. 1). These two bumps represent the time-dependent boundary between the flow and structure. For our numerical experiments the following data setting was used: magnitude of the inlet velocity $v_{in} = 4\,\mathrm{m\,s^{-1}}$, the viscosity $\mu = 15 \cdot 10^{-6}\,\mathrm{kg\,m^{-1}\,s^{-1}}$, the inlet density $\rho_{in} = 1.225\,\mathrm{kg\,m^{-3}}$, the outlet pressure $p_{out} = 97,611$ Pa, the Reynolds number $Re = \rho_{in} v_{in} H/\mu = 5,227$, heat conduction coefficient $k = 2.428 \cdot 10^{-2}\,\mathrm{kg\,m\,s^{-2}\,K^{-1}}$, the specific heat

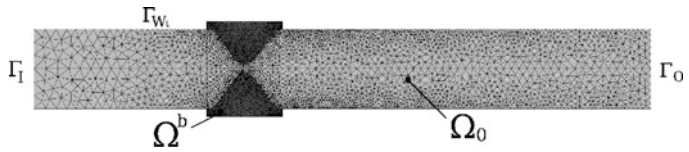


Fig. 1 Computational domain at $t = 0$ with a finite element mesh

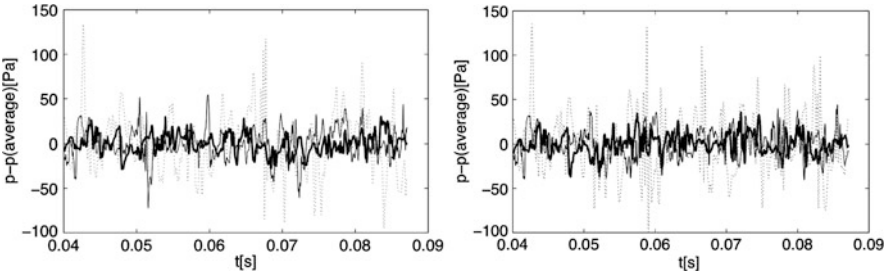


Fig. 2 Dependence of the pressure averaged over the outlet computed on three meshes: strong coupling (*left*), weak coupling (*right*)

$c_v = 721.428\,\mathrm{m^2\,s^{-2}\,K^{-1}}$, the Poisson adiabatic constant $\gamma = 1.4$. The inlet Mach number is $M_{in} = 0.012$. The Young modulus and the Poisson ratio have values $E^b = 25,000$ Pa and $\sigma^b = 0.4$, respectively, the structural damping coefficient is equal to the constant $C = 100\,\mathrm{s^{-1}}$ and the material density $\rho^b = 1,040\,\mathrm{kg\,m^{-3}}$.

Quadratic elements were used for the approximation of the flow problem. For the elasticity problem only linear elements were applied. In Fig. 1 we present the computational mesh with 5398 elements in the flow part and 1998 elements in the structure part (dotted). Along with this computational mesh, two further meshes with different numbers of elements, 10130/2806 (solid) and 20484/4076 (solid thick) elements in the flow/structure part, were used. It allows us to compare the obtained results and to test the influence of the density of the computational meshes on the oscillations of the pressure averaged over the outlet Γ_0. Figure 2 shows the convergence tendency manifested by the decrease of the magnitude of the fluctuations of the quantity p_{av} defined as the pressure averaged over the outlet Γ_O.

In Fig. 2 we cannot observe any significant differences between the results obtained by the weak and strong coupling. This is also well seen from Fig. 3 showing the comparison of the strong and weak coupling on the finest mesh (20484/4076 elements). The main difference is higher stability of strong coupling during the solution of the problem on a long time interval. On the other hand, strong coupling requires longer CPU time.

The second numerical example was performed on a more complicated computational domain. The elastic structure has different parameters in different regions, as seen in Fig. 4. The flow data are the same as above.

Figure 5 shows the flow velocity and the deformation of the computational domain at several time instants during the aeroelastic instability onset.

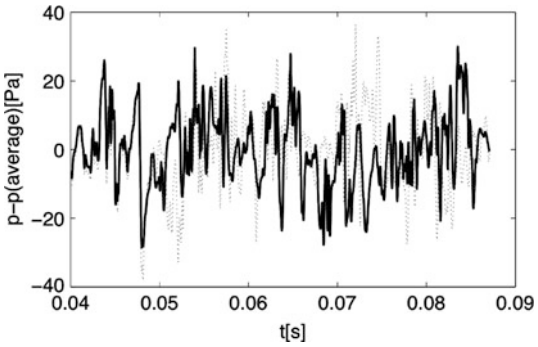


Fig. 3 Comparison of weak (*dotted*) and strong (*strong thick*) coupling for the mesh with 20,484/4,076 elements

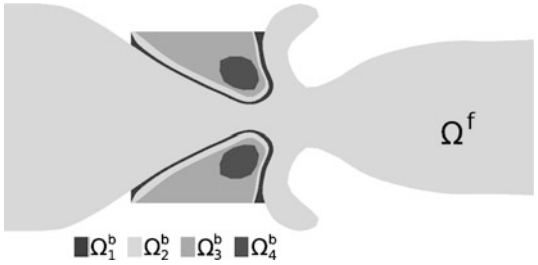


Fig. 4 Scheme of the vocal tract: The material parameters are $E^b = 100$ kPa, $\sigma^b = 0.4$ in Ω_1^b, $E^b = 1$ kPa, $\sigma^b = 0.495$ in Ω_2^b, $E^b = 8$ kPa, $\sigma^b = 0.4$ in Ω_3^b, $E^b = 12$ kPa, $\sigma^b = 0.4$ in Ω_4^b

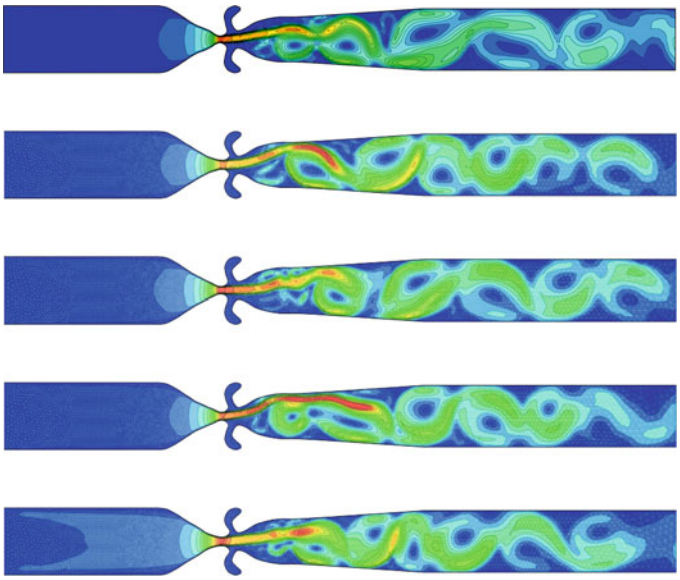


Fig. 5 Velocity isolines at several time instants ($t = 0.261\ s, 0.272\ s, 0.283\ s, 0.294\ s, 0.304\ s$) – $v_{min} = 2\ ms^{-1}, v_{max} = 36\ ms^{-1}$

6 Conclusion

A robust higher-order method for the numerical simulation of the interaction of compressible flow with elastic structures has been presented. The numerical test show the good applicability of the method for the numerical solution of the simplified models of phonation onset. Unfortunately, we are not able to treat the complete closure of the channel by reason of the degeneration of the computational mesh. This remains as the next step of our implementation. Further, the future work will be focused on the deeper analysis of the accuracy of the method and its robustness with respect to the Mach number and Reynolds number, investigation of various types of boundary conditions, solution of more realistic examples and identification of the acoustic signal.

Acknowledgements This work was supported by the grant No. 201/08/0012 (M. Feistauer, V. Kučera) of the Czech Science Foundation, and by the grant SVV-2012-265316 financed by the Charles University in Prague (J. Hasnedlová-Prokopová and A. Kosík).

References

1. Badia, S., Codina, R.: On some fluid-structure iterative algorithms using pressure segregation methods. Application to aeroelasticity. Int. J. Numer. Meth. Engng. **72**, 46–71 (2007)
2. Feistauer, M., Česenek, J., Horáček, J., Kučera, V., Prokopová, J.: DGFEM for the numerical solution of compressible flow in time dependent domains and applications to fluid-structure interaction. In: ECCOMAS CFD 2010, J.C.F. Pereira and A.Sequeira (eds.), Lisbon, Portugal, 14–17 June 2010, CDROM, ISBN 978-989-96778-1-4
3. Feistauer, M., Horáček, J., Kučera, V., Prokopová, J.: On numerical solution of compressible flow in time-dependent domains. Mathematica Bohemica **137**, 1–16 (2012)
4. Feistauer, M., Kučera, V., Prokopová, J.: Discontinuous Galerkin solution of compressible flow in time-dependent domains. Mathematics and Computers in Simulations **80**, 1612–1623 (2010)
5. Kosík, A., Feistauer, M., Horáček, J., Sváček, P.: Numerical simulation of interaction of human vocal folds and fluid flow. In: Náprstek, J., Horáček, J., Okrouhlík, M., Marvalová, B., Verhulst, F., Sawicki, J.T. (eds.) Vibration problems ICOVP 2011, pp. 765–771, Springer Proceedings in Physics 139 (2011)
6. Nomura, T., Hughes, T.J.R.: An arbitrary Lagrangian-Eulerian finite element method for interaction of fluid and a rigid body. Comput. Methods Appl. Mech. Engrg. **95**, 115–138 (1992)

On ε-Uniform Error Estimates For Singularly Perturbed Problems in the DG Method

V. Kučera

Abstract In this paper we present the analysis of the discontinuous Galerkin (DG) finite element method applied to a nonstationary nonlinear convection-diffusion problem. Using the technique of Zhang and Shu (SIAM J Numer Anal 42(2):641–666, 2004), originally for explicit schemes, we prove apriori error estimates uniform with respect to the diffusion coefficient and valid even in the purely convective case. We extend the cited analysis to the method of lines using continuous mathematical induction and a nonlinear Gronwall-type lemma. For an implicit scheme, we prove that there does not exist a Gronwall-type lemma capable of proving the desired estimates using standard arguments. Next, we use a suitable continuation of the implicit solution and use continuous mathematical induction to prove error estimates under a CFL-like condition.

1 Continuous Problem

Let $\Omega \subset \mathbb{R}^d, d = 1, 2, 3$ be a bounded open (polyhedral) domain. We treat the following nonlinear convection-diffusion problem: find $u : \Omega \times (0, T) \to \mathbb{R}$ such that

$$\text{(a)} \quad \frac{\partial u}{\partial t} + \operatorname{div} \mathbf{f}(u) = \varepsilon \Delta u + g \quad \text{in } \Omega \times (0, T), \tag{1}$$

$$\text{(b)} \quad u\big|_{\Gamma_D \times (0,T)} = u_D, \quad \varepsilon \frac{\partial u}{\partial n}\Big|_{\Gamma_N \times (0,T)} = g_N, \tag{2}$$

V. Kučera (✉)
Faculty of Mathematics and Physics, Charles University, Sokolovská 83, 186 75 Praha 8, Prague, Czech Republic
e-mail: vaclav.kucera@email.cz

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_40,

along with the initial condition $u(x,0) = u^0(x)$ in Ω. The diffusion coefficient $\varepsilon \geq 0$ is a given constant, g, u_D, g_N, and u^0 are given functions.

We assume that the convective fluxes $\mathbf{f} = (f_1, \cdots, f_d) \in (C_b^2(\mathbb{R}))^d = (C^2(\mathbb{R}) \cap W^{2,\infty}(\mathbb{R}))^d$, hence $\mathbf{f}$ and $\mathbf{f}' = (f_1', \cdots, f_d')$ are *globally Lipschitz continuous*. For improved estimates via Remark 1, we shall assume $\mathbf{f} \in (C_b^3(\mathbb{R}))^d$. In [4], the error analysis is extended, assuming only local properties, i.e. $\mathbf{f} \in (C^2(\mathbb{R}))^d$ and $\mathbf{f} \in (C^3(\mathbb{R}))^d$.

In our analysis, we need to assume that Γ_N is an outflow boundary for either u or u_h, i.e. e.g. for u, we assume $\Gamma_N^{(t)} \subseteq \{x \in \partial\Omega; \mathbf{f}'(u(x,t)).\mathbf{n} \geq 0\}$ and $\Gamma_D^{(t)} := \partial\Omega \setminus \Gamma_N$.

2 Discretization

Let $\mathscr{T}_h$ be (generally nonconforming) triangulation of $\overline{\Omega}$. For $K \in \mathscr{T}_h$ we set $h = \max_{K\in\mathscr{T}_h} \mathrm{diam}(K)$. By $\mathscr{F}_h$ we denote the system of all faces of all elements $K \in \mathscr{T}_h$. By $\mathscr{F}_h^I, \mathscr{F}_h^D, \mathscr{F}_h^N, \mathscr{F}_h^B$ we denote the sets on interior, Dirichlet, Neumann and boundary edges, respectively. For each $\Gamma \in \mathscr{F}_h$ we define a fixed unit normal $\mathbf{n}_\Gamma$, which has the same orientation as the outer normal to $\partial\Omega$ if $\Gamma \in \mathscr{F}_h^B$.

Over a triangulation $\mathscr{T}_h$ we define the *broken Sobolev spaces* $H^k(\Omega, \mathscr{T}_h) = \{v;\ v|_K \in H^k(K),\ \forall K \in \mathscr{T}_h\}$. For $\Gamma \in \mathscr{F}_h^I$ we have two neighbours $K_\Gamma^{(L)}$, $K_\Gamma^{(R)} \in \mathscr{T}_h$, where $\mathbf{n}_\Gamma$ is the outer normal to $K_\Gamma^{(L)}$. For $v \in H^1(\Omega, \mathscr{T}_h)$ we define on $\Gamma \in \mathscr{F}_h^I$: $v|_\Gamma^{(L)} =$ the trace of $v|_{K_\Gamma^{(L)}}$ on Γ, $v|_\Gamma^{(R)} =$ the trace of $v|_{K_\Gamma^{(R)}}$ on Γ, $\langle v\rangle_\Gamma = \frac{1}{2}\big(v|_\Gamma^{(L)} + v|_\Gamma^{(R)}\big)$ and $[v]_\Gamma = v|_\Gamma^{(L)} - v|_\Gamma^{(R)}$. On $\Gamma \in \mathscr{F}_h^B$ we set $v_\Gamma = v|_\Gamma^{(L)} =$ the trace of $v|_{K_\Gamma^{(L)}}$ on Γ, while $v|_\Gamma^{(R)} = u_D$ on Γ_D, $v|_\Gamma^{(R)} = v|_\Gamma^{(L)}$ on Γ_N.

Let $p \geq 1$ be an integer. The approximate solution will be sought in the space of discontinuous piecewise polynomial functions $S_h = \{v;\ v|_K \in P^p(K), \forall K \in \mathscr{T}_h\}$, where $P^p(K)$ are polynomials on K of degree $\leq p$. By $(\cdot,\cdot)$ we denote the $L^2(\Omega)$-scalar product and by $\|\cdot\|$ the $L^2(\Omega)$-norm. By $\|\cdot\|_\infty$, we denote the $L^\infty(\Omega)$-norm.

We introduce the following forms defined for $v, \varphi \in H^2(\Omega, \mathscr{T}_h)$. *Diffusion form*:

$$a_h(v,\varphi) = \sum_{K\in\mathscr{T}_h} \int_K \nabla v\cdot\nabla\varphi \,\mathrm{d}x - \int_{\mathscr{F}_h^I} \langle\nabla v\rangle\cdot\mathbf{n}[\varphi]\,\mathrm{d}S - \Theta\int_{\mathscr{F}_h^I} \langle\nabla\varphi\rangle\cdot\mathbf{n}[v]\,\mathrm{d}S$$
$$- \int_{\mathscr{F}_h^D} \nabla v\cdot\mathbf{n}\varphi\,\mathrm{d}S - \Theta\int_{\mathscr{F}_h^D} \nabla\varphi\cdot\mathbf{n}v\,\mathrm{d}S.$$

Interior and boundary penalty jump terms:

$$J_h(v,\varphi) = \int_{\mathscr{F}_h^I} \sigma[v][\varphi]\,\mathrm{d}S + \int_{\mathscr{F}_h^D} \sigma v\varphi\,\mathrm{d}S.$$

Right-hand side form:

$$l_h(\varphi)(t) = \int_\Omega g(t)\varphi \,\mathrm{d}x + \int_{\mathscr{F}_h^N} g_N(t)\varphi \,\mathrm{d}S - \varepsilon\Theta \int_{\mathscr{F}_h^D} \nabla\varphi\cdot\mathbf{n}u_D(t)\,\mathrm{d}S + \varepsilon\int_{\mathscr{F}_h^D} \sigma u_D(t)\varphi\,\mathrm{d}S.$$

The parameter σ in the diffusion and right-hand side forms is defined by $\sigma|_\Gamma = C_W|\Gamma|^{-1}$, where $C_W > 0$ is a constant, which is chosen large enough to ensure coercivity of the diffusion form – cf. Lemma 2. Depending on the value of Θ in the diffusion form, we get the *symmetric* ($\Theta = 1$), *incomplete* ($\Theta = 0$) and *nonsymmetric interior penalty* ($\Theta = -1$) variants of the diffusion a right-hand side forms.

Finally we define the *convective form*

$$b_h(v,\varphi) = -\sum_{K\in\mathscr{T}_h}\int_K \mathbf{f}(v)\cdot\nabla v\,\mathrm{d}x + \int_{\mathscr{F}_h^I} H(v^{(L)}, v^{(R)}, \mathbf{n})[\varphi]\,\mathrm{d}S + \int_{\mathscr{F}_h^B} H(v^{(L)}, v^{(R)}, \mathbf{n})\varphi^{(L)}\,\mathrm{d}S.$$

The form b_h approximates convective terms with the aid of a numerical flux $H(v,w,\mathbf{n})$. We assume that H has the following standard properties: H is *Lipschitz-continuous, consistent, conservative* and H is an *E-flux*, i.e.

$$\big(H(v,w,\mathbf{n}) - \mathbf{f}(q)\cdot\mathbf{n}\big)(v-w) \geq 0, \quad \forall v,\ w\in\mathbb{R},\ \mathbf{n}\in B_1 \text{ and all } q \text{ between } v, w.$$

The E-flux condition was introduced as a generalization of monotone fluxes by Osher in [5]. Many numerical fluxes used in practice are E-fluxes, e.g. Lax-Friedrichs, Godunov, Engquist-Osher and the Roe flux with entropy fix, cf. [5].

Definition 1. We say that $u_h \in C^1([0,T]; S_h)$ is a DG solution of (1) and (2), if $u_h(0) = u_h^0 \approx u^0$ and for all $\varphi_h \in S_h$, and $t\in(0,T)$

$$\frac{d}{dt}\big(u_h(t),\varphi_h\big) + b_h\big(u_h(t),\varphi_h\big) + \varepsilon J_h\big(u_h(t),\varphi_h\big) + \varepsilon a_h\big(u_h(t),\varphi_h\big) = l_h\big(\varphi_h\big)(t). \quad (3)$$

3 Some Necessary Results

We assume that the weak solution u is sufficiently regular, namely $u_t := \frac{\partial u}{\partial t} \in L^2\big(0,T; H^{p+1}(\Omega)\big)$, $u \in L^\infty(0,T; W^{1,\infty}(\Omega))$, where $p \geq 1$ is the degree of approximation. These conditions imply $u \in C\big([0,T]; H^{p+1}(\Omega)\big)$.

As for the mesh assumptions, we consider a system $\{\mathscr{T}_h\}_{h\in(0,h_0)}$, $h_0 > 0$, of triangulations, which are shape regular and satisfy the inverse assumption, cf. [2].

Now, for $v\in L^2(\Omega)$ we denote by $\Pi_h v$ the $L^2(\Omega)$-projection of v on S_h:

$$\Pi_h v \in S_h, \quad \big(\Pi_h v - v,\ \varphi_h\big) = 0, \qquad \forall\,\varphi_h \in S_h.$$

Let $\eta_h(t) = u(t) - \Pi_h u(t) \in H^{p+1}(\Omega, \mathscr{T}_h)$ and $\xi_h(t) = \Pi_h u(t) - u_h(t) \in S_h$ for $t \in (0, T)$. Then we can write the error e_h as $e_h(t) := u(t) - u_h(t) = \eta_h(t) + \xi_h(t)$. Standard approximation results give us estimates for $\eta_h(t)$ in terms of power of h, e.g. $||\eta||_{L^2(\Omega)} \leq C h^{p+1} |u|_{H^{p+1}}$, cf. [2].

Lemma 1. *There exists a constant $C \geq 0$ independent of h, t, such that*

$$b_h\big(u_h(t), \xi_h(t)\big) - b_h\big(u(t), \xi_h(t)\big) \leq C\Big(1 + \frac{\|e_h(t)\|_\infty^2}{h^2}\Big)\big(h^{2p+1}|u(t)|_{H^{p+1}}^2 + \|\xi_h(t)\|^2\big).$$

Proof. The proof follows the arguments of [7], where similar estimates are derived for periodic boundary conditions or compactly supported solutions. The proof for mixed Dirichlet-Neumann boundary conditions is contained in [4]. Here, we only note that the estimate is based on performing second order Taylor expansions of and using the *E-flux* properties for H. □

Remark 1. We can improve Lemma (1), if we suppose $\mathbf{f} \in (C_b^3(\mathbb{R}))^d$ and $\Gamma_N = \emptyset$. Then we obtain a factor of $h^{-1}\|e_h\|_\infty^2$ instead of $h^{-2}\|e_h\|_\infty^2$ in the estimate of Lemma (1). This improved estimate will be useful in proving the resulting estimates for lower order polynomials and with a less restrictive CFL condition, cf. Remark 3.

Lemma 2 (Ellipticity and boundedness of A_h, cf. [3]). *Let the constant C_W be large enough. Then the form A_h is* elliptic *and* bounded*, i.e.*

$$\|v\|_{DG}^2 \leq A_h(v, v), \quad \forall v \in H^2(\Omega, \mathscr{T}_h),$$
$$A_h(v, w) \leq \|v\|_{DG}\|w\|_{DG}, \quad \forall v, w \in H^2(\Omega, \mathscr{T}_h),$$

where $\|w\|_{DG}^2 = \frac{1}{2}\big(\sum_{K\in\mathscr{T}_h} |w|_{H^k(K)}^2 + J_h(w, w)\big)$ *and* $A_h(\cdot, \cdot) = a_h(\cdot, \cdot) + J_h(\cdot, \cdot)$.

4 Error Analysis for the Method of Lines

We proceed in a standard way. Due to Galerkin orthogonality, we subtract the equations for u and u_h and set $\varphi_h := \xi_h(t) \in S_h$. Since $\big(\frac{\partial \xi_h}{\partial t}, \xi_h\big) = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\xi_h\|^2$, we get

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\xi_h(t)\|^2 + \varepsilon A_h\big(\xi_h(t), \xi_h(t)\big)$$
$$= -\varepsilon A_h\big(\eta_h(t), \xi_h(t)\big) + b_h\big(u_h(t), \xi_h(t)\big) - b_h\big(u(t), \xi_h(t)\big) - \Big(\frac{\partial \eta_h(t)}{\partial t}, \xi_h(t)\Big).$$

For the last right-hand side term, we use the Cauchy and Young's inequalities and standard estimates for η. For the convective and diffusion terms we use Lemmas 1 and 2. Integration from 0 to $t \in [0, T]$ yields

$$\|\xi_h(t)\|^2 + \int_0^t \varepsilon \|\xi_h(\vartheta)\|_{DG}^2 \,\mathrm{d}\vartheta \tag{4}$$

$$\leq C \int_0^t \Big(1 + \frac{\|e_h(\vartheta)\|_\infty^2}{h^2}\Big)\Big(\big(h^{2p+1} + \varepsilon h^{2p}\big)|u(\vartheta)|_{H^{p+1}}^2 + h^{2p+2}|u_t(\vartheta)|_{H^{p+1}}^2 + \|\xi_h(\vartheta)\|^2\Big)\mathrm{d}\vartheta.$$

For simplicity we have assumed $\xi_h(0) = 0$, i.e. $u_h^0 = \Pi_h u^0$. Otherwise, we must assume e.g. $\|\xi_h(0)\| = O(h^{p+1/2})$ and include this term in (4). We notice that if we knew apriori that $\|e_h\|_\infty = O(h)$ then the unpleasant term $h^{-2}\|e_h\|_\infty^2$ in (4) would be $O(1)$. Thus we could simply apply the standard Gronwall inequality to obtain the desired error estimates.

Lemma 3. *Let $t \in [0, T]$ and $p \geq d/2$. If $\|e_h(\vartheta)\| \leq h^{1+d/2}$ for all $\vartheta \in [0, t]$, then there exists a constant C_T independent of h, t and ε such that*

$$\max_{\vartheta\in[0,t]} \|e_h(\vartheta)\|^2 + \int_0^t \varepsilon \|e_h(\vartheta)\|_{DG}^2 \,\mathrm{d}\vartheta \leq C_T^2\big(h^{2p+1} + \varepsilon h^{2p}\big). \tag{5}$$

Proof. The assumptions imply, using the inverse inequality and estimates of η, that

$$\|e_h(\vartheta)\|_\infty \leq \|\eta_h(\vartheta)\|_\infty + \|\xi_h(\vartheta)\|_\infty \leq Ch|u(t)|_{W^{1,\infty}(\Omega)} + C_I h^{-d/2}\|\xi_h(\vartheta)\|$$

$$\leq Ch + C_I h^{-d/2}\|e_h(\vartheta)\| + C_I h^{-d/2}\|\eta_h(\vartheta)\| \leq Ch + Ch^{p+1-d/2}|u(\vartheta)|_{H^{p+1}} \leq Ch,$$

where the constant C is independent of h, ϑ, t. Using this estimate in (4) gives us

$$\|\xi_h(t)\|^2 + \int_0^t \varepsilon \|\xi_h(\vartheta)\|_{DG}^2 \,\mathrm{d}\vartheta \leq \tilde{C}\big(h^{2p+1} + \varepsilon h^{2p}\big) + C\int_0^t \|\xi_h(\vartheta)\|^2 \,\mathrm{d}\vartheta,$$

Applying Gronwall's inequality gives us the desired estimate for ξ_h, which along with similar estimates for η gives us (5). □

Now it remains to get rid of the *apriori* assumption $\|e_h\|_\infty = O(h)$. In [7] this is done for an explicit scheme using mathematical induction. Starting from $\|e_h^0\| = O(h^{p+1/2})$, the following induction step is proved:

$$\|e_h^n\| = O(h^{p+1/2}) \quad \Longrightarrow \quad \|e_h^{n+1}\|_\infty = O(h) \quad \Longrightarrow \quad \|e_h^{n+1}\| = O(h^{p+1/2}). \tag{6}$$

For the method of lines we have no discrete structure with respect to time and hence cannot use mathematical induction straightforwardly. However, we can divide $[0, T]$ into a finite number of sufficiently small intervals $[t_n, t_{n+1}]$ on which "e_h *does not change too much*" and use induction with respect to n. This is essentially a *continuous mathematical induction* argument, a concept introduced in [1].

Remark 2. Due to the regularity assumptions, $u, u_h \in C([0, T]; L^2(\Omega))$. Since $[0, T]$ is a compact set, $e_h(\cdot)$ is a *uniformly continuous* function from $[0, T]$ to $L^2(\Omega)$, i.e.

$$\forall \bar{\epsilon} > 0\ \exists \delta > 0 :\ s, \bar{s} \in [0, T], |s - \bar{s}| \le \delta \Rightarrow \|e_h(s) - e_h(\bar{s})\| \le \bar{\epsilon}.$$

Theorem 1 (Main theorem). *Let $p > 1 + d/2$. Then there exists $h_1 > 0$ such that for all $h \in (0, h_1]$ we have the estimate*

$$\max_{\vartheta \in [0,T]} \|e_h(\vartheta)\|^2 + \int_0^T \varepsilon \|e_h(\vartheta)\|_{DG}^2 \, \mathrm{d}\vartheta \le C_T^2 \big(h^{2p+1} + \varepsilon h^{2p}\big).$$

Proof. We have $p > 1+d/2$, thus for all sufficiently small h, we have $C_T(h^{p+1/2} + \sqrt{\varepsilon} h^p) \le \frac{1}{2} h^{1+d/2}$. We fix an arbitrary h. By Remark 2, there exists $\delta > 0$, such that if $s, \bar{s} \in [0, T], |s - \bar{s}| \le \delta$, then $\|e_h(s) - e_h(\bar{s})\| \le \frac{1}{2} h^{1+d/2}$. We define $t_i = i\delta$, $i = 0, 1, \ldots$ and set $N := \max\{i = 0, 1, \ldots; t_i < T\}$, $t_{N+1} := T$. This defines a partition $0 = t_0 < t_1 < \cdots < t_{N+1} = T$ of $[0, T]$ into $N + 1$ intervals of length (at most) δ.

We shall now prove by induction that for all $n = 1, \ldots, N + 1$

$$\max_{\vartheta \in [0,t_n]} \|e_h(\vartheta)\|^2 + \int_0^{t_n} \varepsilon \|e_h(\vartheta)\|_{DG}^2 \, \mathrm{d}\vartheta \le C_T^2 \big(h^{2p+1} + \varepsilon h^{2p}\big). \tag{7}$$

The desired error estimate is thus obtained by taking $n := N + 1$ in (7).

(i) $n = 1$: Since $\|e_h(0)\| \le \frac{1}{2} h^{1+d/2}$. By uniform continuity, we have for all $s \in [0, t_1]$

$$\|e_h(s)\| \le \|e_h(0)\| + \|e_h(s) - e_h(0)\| \le \tfrac{1}{2} h^{1+d/2} + \tfrac{1}{2} h^{1+d/2} = h^{1+d/2}.$$

Therefore, by Lemma 3 we obtain estimate (7) on $[0, t_1]$, i.e. for $n = 1$.

(ii) Induction step: We assume that (7) holds for general $n < N + 1$. Therefore $\|e_h(t_n)\| \le C_T(h^{p+1/2} + \sqrt{\varepsilon} h^p) \le \frac{1}{2} h^{1+d/2}$. By uniform continuity, for all $s \in [t_n, t_{n+1}]$

$$\|e_h(s)\| \le \|e_h(t_n)\| + \|e_h(s) - e_h(t_n)\| \le \tfrac{1}{2} h^{1+d/2} + \tfrac{1}{2} h^{1+d/2} = h^{1+d/2}.$$

This and the induction assumption imply that $\|e_h(s)\| \le h^{1+d/2}$ for $s \in [0, t_n] \cup [t_n, t_{n+1}] = [0, t_{n+1}]$. By Lemma 3, we obtain estimate (7) on $[0, t_{n+1}]$. □

Remark 3. If we assume $\mathbf{f} \in (C_b^3(\mathbb{R}))^d$ then by Remark 1 we get the improved assumption $p > (1+d)/2$ in Theorem 1. If $\varepsilon = 0$ we need to assume only $p > d/2$.

Remark 4. For the method of lines we can use a *nonlinear Gronwall-type lemma* to prove Theorem 1 directly, cf. [4]. As stated in Remark 6, this is not possible for an implicit scheme, since an analogous discrete Gronwall lemma cannot exist.

5 Error Estimates for a Fully Implicit Scheme

In this section, we shall introduce and analyze the DG scheme with a standard implicit Euler time discretization. Here we cannot use the approach of [7] for the explicit scheme, since we were unable to prove the first implication in the induction step (6). On the other hand, in Lemma 6 we prove that for the implicit Euler scheme we cannot use a discrete Gronwall-type lemma as mentioned in Remark 4.

We consider a partition $0 = t_0 < t_1 < \cdots < t_{N+1} = T$ of $[0, T]$ and set $\tau_n = t_{n+1} - t_n$ for $n = 0, \cdots, N$. The exact solution $u(t_n)$ will be approximated by $u_h^n \in S_h$.

Definition 2. We say that $\{u_h^n\}_{n=0}^N \subset S_h$ is an implicit Euler DGFE solution of the convection-diffusion problem (1) and (2), if $u_h^0 = \Pi_h u^0$ and for all $\varphi_h \in S_h, n = 0, \cdots, N$

$$\Big(\frac{u_h^{n+1} - u_h^n}{\tau_n}, \varphi_h\Big) + b_h\big(u_h^{n+1}, \varphi_h\big) + \varepsilon A_h\big(u_h^{n+1}, \varphi_h\big) = l_h\big(\varphi_h\big)(t_{n+1}). \tag{8}$$

Similarly as in Sect. 3, we define $\eta_h^n = u(t_n) - \Pi_h u(t_n) \in H^{p+1}(\Omega, \mathscr{T}_h)$ and $\xi_h^n = \Pi_h u(t_n) - u_h^n \in S_h$. Then we can write the error e_h^n as $e_h^n := u(t_n) - u_h^n = \eta_h^n + \xi_h^n$.

First, we analyze problem (8), proving that u_h^{n+1} exists uniquely and depends continuously on τ_n. To this end we define an abstract formulation of problem (8):

Definition 3. (Auxiliary problem) Let $t \in [0, T]$, $\tau \in [0, T]$ and $U_h \in S_h$. We seek $u_\tau \in S_h$ such that

$$\big(u_\tau - U_h, \varphi_h\big) + \tau b_h\big(u_\tau, \varphi_h\big) + \tau\varepsilon A_h\big(u_\tau, \varphi_h\big) = \tau l_h\big(\varphi_h\big)(t), \quad \forall \varphi_h \in S_h. \tag{9}$$

Remark 5. If we take $\tau := \tau_n, U_h := u_h^n, t := t_{n+1}$ and define $u_h^{n+1} := u_\tau$, the auxiliary problem (9) reduces to equation (8), which defines u_h^{n+1}. If we take $\tau := 0$ the solution of (9) is $u_\tau = u_h^n$. Between these two cases u_τ depends continuously on τ:

Lemma 4. *There exist constants $C_1, C_2 > 0$ independent of h, τ, t, ε, such that the following holds. Let $t \in [0, T], h \in (0, h_0), U_h \in S_h$ and $\tau \in [0, \tau_0)$, where $\tau_0 = \max\{C_1\varepsilon, C_2 h\}$. Then the solution u_τ of (9) exists, is uniquely determined and $\|u_\tau\|$ depends continuously on $\tau \in [0, \tau_0)$.*

Proof. Problem (9) is a nonlinear equation for u_τ on the finite-dimensional space S_h. The statements follow from the nonlinear Lax-Milgram theorem, cf. [6]. For details of the proof, see [4]. □

Definition 4 (Continuated discrete solution). Let $\tilde{u}_h : [0, T] \to S_h$ such that for $s \in [t_n, t_{n+1}]$ we set $\tilde{u}_h(s) := u_\tau$, the solution of the auxiliary problem (9) with $\tau := s - t_n$, t $:= t_{n+1}$ and $U_h := u_h^n$. Furthermore, we define $\tilde{e}_h := u - \tilde{u}_h$ and $\tilde{\xi}_h := \Pi_h u - \tilde{u}_h$.

Under the assumptions of Lemma 4, $\tilde{u}_h, \tilde{e}_h \in C([0,T];L^2(\Omega))$ and $\tilde{u}_h$ is uniquely determined. Also, $\tilde{u}_h(t_n) = u_h^n$ and $\tilde{e}_h(t_n) = e_h^n$ for $n = 0, \cdots, N$. Therefore, estimates of $\tilde{e}_h(\cdot)$ imply estimates of e_h^n. Since $\tilde{u}_h$ is constructed using problem (9), which is essentially the implicit scheme (8) with special data, we can derive error estimates for $\tilde{u}_h$ in a standard manner. For simplicity we assume a uniform partition of $[0,T]$.

Lemma 5. *Let $p > d/2$ and $s \in (t_n, t_{n+1}]$ for some $n \in \{0, \cdots, N-1\}$. If $\|\tilde{e}_h(s)\| \le h^{1+d/2}$ and $\|\tilde{e}_h(t_k)\| \le h^{1+d/2}$ for all $k = 0, \cdots, n$, then there exists $C_T > 0$ independent of s, n, h, τ such that*

$$\max_{t\in\{t_0,\cdots,t_n,s\}} \|\tilde{e}_h(t)\|^2 + \sum_{k=1}^{n} \tau\varepsilon\|\tilde{e}_h(t_k)\|_{DG}^2 + (s-t_n)\varepsilon\|\tilde{e}_h(s)\|_{DG}^2 \le C_T^2\big(h^{2p+1} + \varepsilon h^{2p} + \tau^2\big).$$

Proof. We subtract (9) from the equation for the exact solution. Thus $\tilde{e}_h(s)$ satisfies

$$\begin{aligned}\big(\tilde{e}_h(s) - \tilde{e}_h(t_n), \varphi_h\big) + (s-t_n)\big(b_h(u(s),\varphi_h) - b_h(\tilde{u}_h(s),\varphi_h)\big) + (s-t_n)\varepsilon A_h(\tilde{e}_h(s),\varphi_h)\\ = \big(u(s) - u(t_n) - (s-t_n)u_t(s), \varphi_h\big). \qquad (10)\end{aligned}$$

We set $\varphi_h := \tilde{\xi}_h(s)$ and use the fact that $2(a-b,a) = \|a\|^2 - \|b\|^2 + \|a-b\|^2$. We estimate the convective terms using Lemma 1 and the diffusion terms using Lemma 2. The right-hand side represents the temporal error and is estimated as usual. Thus

$$\begin{aligned}&\|\tilde{\xi}_h(s)\|^2 - \|\tilde{\xi}_h(t_n)\|^2 + \|\tilde{\xi}_h(s) - \tilde{\xi}_h(t_n)\|^2 + (s-t_n)\varepsilon\|\tilde{\xi}_h(s)\|_{DG}^2\\ &\le C\tau\Big(1 + \frac{\|\tilde{e}_h(s)\|_\infty^2}{h^2}\Big)\Big((h^{2p+1} + \varepsilon h^{2p})|u|_{L^\infty(H^{p+1})}^2 + \tau^2\|u_{tt}\|_{L^\infty(L^2(\Omega))}^2 + \|\tilde{\xi}_h(s)\|^2\Big).\end{aligned}$$

The assumptions imply $\|\tilde{e}_h(s)\|_\infty \le Ch$, eliminating the factor h^{-2}. Thus

$$\|\tilde{\xi}_h(s)\|^2 + (s-t_n)\varepsilon\|\tilde{\xi}_h(s)\|_{DG}^2 \le \|\tilde{\xi}_h(t_n)\|^2 + C\tau\big(h^{2p+1} + \varepsilon h^{2p} + \tau^2 + \|\tilde{\xi}_h(s)\|^2\big).$$

Similarly, we may derive estimates at t_{k+1}:

$$\|\tilde{\xi}_h(t_{k+1})\|^2 + \tau\varepsilon\|\tilde{\xi}_h(t_{k+1})\|_{DG}^2 \le \|\tilde{\xi}_h(t_k)\|^2 + C\tau\big(h^{2p+1} + \varepsilon h^{2p} + \tau^2 + \|\tilde{\xi}_h(t_{k+1})\|^2\big).$$

Combining these estimates and using the discrete Gronwall lemma gives us the desired estimate for $\tilde{\xi}_h$. Standard estimates for η give us the estimate for $\tilde{e}_h$. □

Theorem 2 (Main theorem – implicit version). *Let $p > 1 + d/2$. Let $h_1, \tau_1 > 0$ be such that $C_T(h_1^{p+1/2} + \sqrt{\varepsilon}h_1^p + \tau_1) = \frac{1}{2}h_1^{1+d/2}$ and $\tau_1 < \tau_0$, where τ_0 is defined in Lemma 4. Then for all $h \in (0, h_1)$, $\tau \in (0, \tau_1)$ we have the estimate*

$$\max_{n\in\{0,\cdots,N\}} \|e_h^n\|^2 + \sum_{n=1}^{N} \tau\left(\varepsilon\|e_h^n\|_{DG}^2 + \widetilde{J}_h\left(e_h^n, e_h^n\right)\right) \le C_T^2\left(h^{2p+1} + \varepsilon h^{2p} + \tau^2\right). \quad (11)$$

Proof. Again, $\tilde{e}_h(\cdot)$ is a *uniformly continuous* function from $[0, T]$ to $L^2(\Omega)$. This allows to use continuous mathematical induction to eliminate the apriori assumption $\|\tilde{e}_h(t)\| = O(h^{1+d/2})$ from Lemma 5. The proof thus follows that of Theorem 1. □

Remark 6. The reason we introduced the continuation of u_h^n is that a more standard, straightforward approach is insufficient. Specifically, we prove in [4] that there does not exist a Gronwall-type lemma which could prove the desired error estimate (11) *only* from the error equation of the implicit scheme tested by ξ_h^{n+1} and the derived estimates of individual terms contained therein.

6 Conclusion

We have presented an analysis of the DG method for a nonlinear convection-diffusion problem. Building on results from [7], which dealt with an explicit time discretization, we proved apriori $L^\infty(L^2)$ error estimates independent of the diffusion coefficient for the method of lines and a fully implicit scheme. We have derived the key estimates for the case of mixed Dirichlet-Neumann boundary conditions, improving the results of [7]. For the method of lines, the error estimates are derived using a continuous mathematical induction argument or a nonlinear Gronwall lemma. For the implicit time discretization, we show that a similar discrete Gronwall lemma does not exist and prove the error estimates using continuous mathematical induction applied to a suitable continuation of the discrete solution. However, using this technique, we obtain an unnatural CFL-like condition for the implicit scheme. In [4], the presented results are extended to of a *locally* Lipschitz continuous **f**.

Acknowledgements The work was supported by the project P201/11/P414 of the Czech Science Foundation.

References

1. Chao Y. R., *A note on "Continuous mathematical induction"*, Bull. Amer. Math. Soc., **26** (1), 17–18 (1919).
2. Ciarlet P.G., *The Finite Element Method for Elliptic Problems,* North-Holland, Amsterdam (1979).
3. Dolejší V., Feistauer M., Kučera V. and Sobotíková V., *An optimal $L^\infty(L^2)$-error estimate for the discontinuous Galerkin approximation of a nonlinear non-stationary convection-diffusion problem*, IMA J. Numer. Anal., **28**, 496–521 (2008).

4. Kučera V., *On diffusion-uniform error estimates for the DG method applied to singularly perturbed problems*, The Preprint Series of the School of Mathematics, preprint No. MATH-knm-2011/3 (2011), http://www.karlin.mff.cuni.cz/ms-preprints/prep.php.
5. Osher S., *Riemann solvers, the entropy condition, and difference approximations*, SIAM. J. Numer. Anal., **21**, 217–235 (1984).
6. Zeidler E., *Nonlinear Functional Analysis and Its Applications II/B: Nonlinear Monotone Operators*, Springer (1986).
7. Zhang Q. and Shu C.-W., *Error estimates to smooth solutions of Runge–Kutta discontinuous Galerkin methods for scalar conservation laws*, SIAM J. Numer. Anal., **42**(2), 641–666 (2004).

Two-Sided a Posteriori Error Estimates for the DGMs for the Heat Equation

I. Šebestová

Abstract We derive a two-sided error bound for the nonstationary heat equation with mixed Dirichlet/Neumann boundary conditions. The space semi-discretization is carried out with the aid of the interior penalty discontinuous Galerkin methods and the backward Euler method is employed for the time discretization. The approach is based on the Helmholtz decomposition and the averaging interpolation operator. The behavior of derived estimates is demonstrated on a numerical example.

1 Introduction

In this paper we are dealing with the a posteriori error analysis for the heat conduction equation accompanied with mixed Dirichlet/Neumann boundary conditions. There is a number of papers devoted to a posteriori error estimates (AEE) for this equation. The paper [4] is based on potential and flux reconstructions together with triangle inequality. The analysis is done for various finite element methods. Taking a suitable norm for measuring the error, the estimates are free of unknown constants. But it requires a reconstruction of the flux from the RTN space. The paper [8] is based on the Helmholtz decomposition of the gradient of the error and it is carried out for Crouzeix-Raviart finite element method. In [10] AEE for classical Galerkin method is derived and classical techniques from [9] are applied. In [5] the upper bound for the problem was derived with the aid of the elliptic reconstruction of the DG-approximation. We have used the Helmholtz decomposition to extend the results from [3] to the high-order discontinuous Galerkin method (DGM) and, moreover, we have derived the lower error bound.

I. Šebestová (✉)
Faculty of Mathematics and Physics, Charles University in Prague, Sokolovská 83, 186 75, Praha 8, Czech Republic
e-mail: ivasebestova@seznam.cz

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_41,

2 Problem Definition and Discretization

We are going to state the continuous problem. Further, we will carry out the space and time discretization with the aid of DGM and the backward Euler scheme, respectively. Finally, notation used throughout the text will be introduced.

Let $\Omega \subset \mathbb{R}^d$ ($d = 2$ or 3) be a bounded multiply connected polyhedral Lipschitz domain with a boundary $\partial\Omega = \partial\Omega_D \cup \partial\Omega_N$, $T > 0$, and $Q_T = \Omega \times (0, T)$. Let us consider the problem:

$$\begin{aligned} \partial u/\partial t - \Delta u &= f \quad && \text{in } Q_T, \\ u &= u_D && \text{on } \partial\Omega_D \times (0,T), \\ \nabla u \cdot n &= g_N && \text{on } \partial\Omega_N \times (0,T), \\ u(x,0) &= u^0(x) && \text{in } \Omega. \end{aligned} \tag{1}$$

We use a standard notation for the function spaces. Moreover, we denote $H^1_D(\Omega) \equiv \{v \in H^1(\Omega);\ v = 0 \quad \text{on} \quad \partial\Omega_D\}$.

2.1 Time and Space Discretization

Let $0 = t_0 < t_1 < \ldots < t_{\bar{N}} = T$ be a partition of the time interval $[0, T]$ and let $\tau_n = t_n - t_{n-1}$, $\tau = \max\{\tau_n : 1 \le n \le \bar{N}\}$. We use the backward Euler scheme to get the *semi-discrete problem*: Find a sequence $\{u^n\}_{1 \le n \le \bar{N}}$, $u^n - u^*(t_n) \in H^1_D(\Omega)$ such that

$$\int_\Omega \frac{u^n - u^{n-1}}{\tau_n} v \,\mathrm{d}\mathbf{x} + \int_\Omega \nabla u^n \cdot \nabla v \,\mathrm{d}\mathbf{x} = \int_\Omega f^n v \,\mathrm{d}\mathbf{x} + \int_{\partial\Omega_N} g_N^n v \,\mathrm{d}S \quad \forall v \in H^1_D(\Omega), \tag{2}$$

where $u^*(t_n) \in H^1(\Omega)$ has the trace $u_D^n := u_D(\cdot, t_n)$ on $\partial\Omega_D$, $f^n := f(\cdot, t_n)$, and $g_N^n := g_N(\cdot, t_n)$. For simplicity, we assume that functions u_D^n, f^n, and g_N^n are piecewise polynomial for each time t_n. The solution of (2) is called the *semi-discrete solution*.

As mentioned before, we will carry out the space discretization with the aid of the high-order DGM. On each time level t_n, $n = 1, \ldots, \bar{N}$, we consider a family $\{\mathcal{T}_h^n\}_{h>0}$ of partitions of the closure of Ω into a finite number of closed simplices with mutually disjoint interiors, possibly containing hanging nodes. These partitions are called triangulations hereafter. We assume that meshes are shape regular, locally quasi-uniform, and that there exists a triangulation $\widetilde{\mathcal{T}}_h^n$, which is a refinement of both $\mathcal{T}_h^{n-1}$ and $\mathcal{T}_h^n$, $1 \le n \le \bar{N}$, satisfying in addition the transition condition, which means a local restriction on the refinement/coarsening between two consecutive time levels.

By $\widetilde{\mathcal{F}}_h^{n,\mathrm{I}}$, $\widetilde{\mathcal{F}}_h^{n,\mathrm{D}}$, and $\widetilde{\mathcal{F}}_h^{n,\mathrm{N}}$ we denote the set of all interior faces, faces on $\partial\Omega_D$, and faces on $\partial\Omega_N$, respectively. For simplicity, we put $\widetilde{\mathcal{F}}_h^{n,\mathrm{ID}} := \widetilde{\mathcal{F}}_h^{n,\mathrm{I}} \cup \widetilde{\mathcal{F}}_h^{n,\mathrm{D}}$, $\widetilde{\mathcal{F}}_h^{n,\mathrm{DN}} := \widetilde{\mathcal{F}}_h^{n,\mathrm{D}} \cup \widetilde{\mathcal{F}}_h^{n,\mathrm{N}}$, and $\widetilde{\mathcal{F}}_h^{n} := \widetilde{\mathcal{F}}_h^{n,\mathrm{I}} \cup \widetilde{\mathcal{F}}_h^{n,\mathrm{D}} \cup \widetilde{\mathcal{F}}_h^{n,\mathrm{N}}$. We set $h_K = \mathrm{diam}(K)$

and ∂K denotes the boundary of element K for $K \in \widetilde{\mathscr{T}}_h^n$. For $\Gamma \in \widetilde{\mathscr{F}}_h^{n,\mathrm{I}}$, we set $h_\Gamma := \max(h_{K_\Gamma^{\mathrm{L}}}, h_{K_\Gamma^{\mathrm{R}}})$, where $\Gamma \subset \overline{K_\Gamma^{\mathrm{L}}} \cap \overline{K_\Gamma^{\mathrm{R}}}$. For $\Gamma \in \widetilde{\mathscr{F}}_h^{n,\mathrm{DN}}$, we put $h_\Gamma := \max h_{K_\Gamma^{\mathrm{L}}}$, $\Gamma \subset \partial K_\Gamma^{\mathrm{L}}$.

We denote $\mathbf{n}_\Gamma$, $[v]_\Gamma$, and $\langle v \rangle_\Gamma$ a unit normal vector, the jump and the average over a face Γ, respectively. If $\mathbf{n}_\Gamma$, $[\cdot]_\Gamma$, and $\langle \cdot \rangle_\Gamma$ appear in an integral of the form $\int_\Gamma \dots dS$, we will omit the subscript Γ. Finally, $H^s(\Omega, \widetilde{\mathscr{T}}_h^n)$ denotes the so-called broken Sobolev space, S_{hp}^n the space of discontinuous piecewise polynomial functions, and Π_{hp} the L^2-projection operator on S_{hp}^n. For $u_h^n, v_h^n \in H^2(\Omega, \widetilde{\mathscr{T}}_h^n)$, we define

$$
\begin{aligned}
a_h^n(u_h^n, v_h^n) := & \sum_{K \in \widetilde{\mathscr{T}}_h^n} \int_K \nabla u_h^n \cdot \nabla v_h^n \, \mathrm{d}\mathbf{x} - \sum_{\Gamma \in \widetilde{\mathscr{F}}_h^{n,\mathrm{ID}}} \int_\Gamma \langle \nabla u_h^n \cdot \mathbf{n} \rangle [v_h^n] \, \mathrm{d}S \\
& + \theta \sum_{\Gamma \in \widetilde{\mathscr{F}}_h^{n,\mathrm{ID}}} \int_\Gamma \langle \nabla v_h^n \cdot \mathbf{n} \rangle [u_h^n] \, \mathrm{d}S + \sum_{\Gamma \in \widetilde{\mathscr{F}}_h^{n,\mathrm{ID}}} \int_\Gamma \sigma [u_h^n][v_h^n] \, \mathrm{d}S, \\
\ell_h^n(v_h^n) := & \int_\Omega f^n v_h^n \, dx + \sum_{\Gamma \in \widetilde{\mathscr{F}}_h^{n,\mathrm{N}}} \int_\Gamma g_N^n v_h^n \, dS + \theta \sum_{\Gamma \in \widetilde{\mathscr{F}}_h^{n,\mathrm{D}}} \int_\Gamma \nabla v_h^n \cdot \mathbf{n} u_D^n \, \mathrm{d}S \\
& + \sum_{\Gamma \in \widetilde{\mathscr{F}}_h^{n,\mathrm{D}}} \int_\Gamma \sigma u_D^n v_h^n \, \mathrm{d}S,
\end{aligned}
$$

where $u_D^n = u_D(\cdot, t_n)$, σ is the penalty parameter, and the parameter $\theta = -1$, $\theta = 1$, and $\theta = 0$ corresponds to the symmetric, nonsymmetric, and incomplete variants of the DGM, respectively.

Now, we can state the discrete problem: For a given approximation $u_h^0 \in S_{hp}^0$ of an initial condition u^0, find a sequence $\{u_h^n\}_{1 \le n \le \bar{N}}$, $u_h^n \in S_{hp}^n$ such that

$$
\int_\Omega \frac{u_h^n - u_h^{n-1}}{\tau_n} v_h^n \, \mathrm{d}\mathbf{x} + a_h^n(u_h^n, v_h^n) = \ell_h^n(v_h^n) \quad \forall v_h^n \in S_{hp}^n. \tag{3}
$$

We call the solution of (3) the approximate solution. The reader is referred to [1] for the derivation of discontinuous Galerkin formulation. Let $\{u^n\}_{1 \le n \le \bar{N}}$ be the semi-discrete solution given by (2) and $\{u_h^n\}_{1 \le n \le \bar{N}}$ be the approximate solution given by (3). We set $\{e^n\}_{1 \le n \le \bar{N}} = \{u^n - u_h^n\}_{1 \le n \le \bar{N}}$.

3 A Posteriori Error Analysis: Upper and Lower Error Bound

In this section, we state two main theorems providing upper and lower error bound. First, we will introduce Helmholtz decomposition and an appropriate interpolation operator, as they form the basis of the presented approach developed in [8]. The

idea of using Helmholtz decomposition for splitting the error into conforming and nonconforming parts goes back to the paper [2]. The error is measured in the norm combining the L^2-norm on the last time level and H^1-seminorm on all time levels (except the initial one). In what concerns lower error estimates, all terms except for the interelement jumps term in local error indicators (8) can be estimated in a standard way using suitable cut-off functions (see [9]). Nevertheless, the interelement jumps term can be estimated easily using only the discrete scheme (3) as have been done in [4]. Hereafter, c denotes a generic positive constant, which can differ from formula to formula and is independent of h and τ.

We are interested in the interpolation operator $I_h^{n,D}$ that maps $H^1(\Omega, \widetilde{\mathscr{T}}_h^n)$ into $S_{hp}^n \cap H_D^1(\Omega)$. Denoting by $\mathscr{I}_{\mathrm{Av}}^{\mathrm{D}}$ the standard averaging operator (see [7]), we define

$$I_h^{n,D}(v) = \mathscr{I}_{\mathrm{Av}}^{\mathrm{D}}(\Pi_{hp}(v)) \quad \forall\, v \in H^1(\Omega, \widetilde{\mathscr{T}}_h^n).$$

The Helmholtz decomposition of the (broken) gradient of the error e^n reads (for the properties of such splitting see, e.g., [2])

$$\nabla_h e^n = \nabla \phi^n + \operatorname{curl} \chi^n. \tag{4}$$

Moreover, the error e^n satisfies the following relations as proved in [6]. Analogous relations for Crouzeix-Raviart finite element method have been proved in [8].

Lemma 1. *Let $v_h \in S_{hp}^n \cap H_D^1(\Omega)$, $\phi \in H_D^1(\Omega)$ and $\chi \in (H^1(\Omega))^k$ ($k = 1$ for $d = 2$ and $k = 3$ for $d = 3$) such that $\mathbf{n} \cdot \operatorname{curl} \chi = 0$ on $\partial\Omega_N$. The error e^n satisfies*

$$\sum_{K \in \widetilde{\mathscr{T}}_h^n} \int_K \nabla e^n \cdot \nabla v_h \,\mathrm{d}\mathbf{x} = \int_\Omega \frac{e^{n-1} - e^n}{\tau_n} v_h \,\mathrm{d}\mathbf{x} + \theta \sum_{\Gamma \in \widetilde{\mathscr{F}}_h^{n,\mathrm{I}}} \int_\Gamma \langle \nabla v_h \cdot \mathbf{n} \rangle [u_h^n] \,\mathrm{d}S, \tag{5}$$

$$\sum_{K \in \widetilde{\mathscr{T}}_h^n} \int_K \nabla e^n \cdot \nabla \phi \,\mathrm{d}\mathbf{x} = \int_\Omega (f^n - \frac{u^n - u^{n-1}}{\tau_n}) \phi \,\mathrm{d}\mathbf{x} - \sum_{K \in \widetilde{\mathscr{T}}_h^n} \int_{\partial K} \nabla u_h^n \cdot \mathbf{n} \phi \,\mathrm{d}S + \sum_{K \in \widetilde{\mathscr{T}}_h^n} \int_K \Delta u_h^n \phi \,\mathrm{d}\mathbf{x} + \int_{\partial\Omega_N} g_N^n \phi \,\mathrm{d}S, \tag{6}$$

$$\sum_{K \in \widetilde{\mathscr{T}}_h^n} \int_K \nabla (e^n - \phi) \cdot \operatorname{curl} \chi \,\mathrm{d}\mathbf{x} = \sum_{K \in \widetilde{\mathscr{T}}_h^n} \int_{\partial K \setminus \partial\Omega_N} (e^n - \phi) \operatorname{curl} \chi \cdot \mathbf{n} \, dS. \tag{7}$$

First, let us introduce some additional notation:

$$\mathrm{R}_K^n := \left(f^n + \Delta u_h^n - \frac{u_h^n - u_h^{n-1}}{\tau_n} \right) |_K, \quad K \in \widetilde{\mathscr{T}}_h^n,$$

$$(\eta_{\mathrm{R}}^n)^2 := \sum_{K \in \widetilde{\mathscr{T}}_{h,n}} h_K^2 \left\| \mathrm{R}_K^n \right\|_K^2, \qquad (\eta_{\mathrm{J}}^n)^2 := \sum_{\Gamma \in \widetilde{\mathscr{F}}_h^{n,\mathrm{I}}} h_\Gamma^{-1} \| [u_h^n] \|_\Gamma^2,$$

$$(\eta_{\mathrm{Jd}}^n)^2 := \sum_{\Gamma \in \widetilde{\mathscr{F}}_h^{n,\mathrm{I}}} h_\Gamma \| [\mathbf{n} \cdot \nabla u_h^n] \|_\Gamma^2, \qquad (\eta_{\mathrm{Jdb}}^n)^2 := \sum_{\Gamma \in \widetilde{\mathscr{F}}_h^{n,\mathrm{N}}} h_\Gamma \left\| g_N^n - \mathbf{n} \cdot \nabla u_h^n \right\|_\Gamma^2.$$

For time level $n \geq 1$ we define the *local error indicators* as follows:

$$\begin{aligned} \eta_{K,1}^n = {} & h_K \left\| \mathrm{R}_K^n \right\|_K + \sum_{\Gamma \in \widetilde{\mathscr{F}}_K^{n,\mathrm{N}}} h_\Gamma^{1/2} \left\| g_N^n - \mathbf{n} \cdot \nabla u_h^n \right\|_\Gamma \\ & + \sum_{\Gamma \in \widetilde{\mathscr{F}}_K^{n,\mathrm{I}}} h_\Gamma^{1/2} \| [\mathbf{n} \cdot \nabla u_h^n] \|_\Gamma + \sum_{\Gamma \in \widetilde{\mathscr{F}}_K^{n,\mathrm{I}}} h_\Gamma^{-1/2} \| [u_h^n] \|_\Gamma + \sum_{\Gamma \in \widetilde{\mathscr{F}}_K^{n,\mathrm{D}}} h_\Gamma^{-1/2} \| u_D^n - u_h^n \|_\Gamma, \\ \eta_{K,2}^n = {} & \sum_{\Gamma \in \widetilde{\mathscr{F}}_K^{n,\mathrm{I}}} h_\Gamma^{1/2} \| [u_h^n] \|_\Gamma + \sum_{\Gamma \in \widetilde{\mathscr{F}}_K^{n,\mathrm{D}}} h_\Gamma^{1/2} \| u_D^n - u_h^n \|_\Gamma, \end{aligned} \tag{8}$$

where $\widetilde{\mathscr{F}}_K^{n,\mathrm{I}}$, $\widetilde{\mathscr{F}}_K^{n,\mathrm{N}}$, $\widetilde{\mathscr{F}}_K^{n,\mathrm{D}}$ denote the set of all interior faces of element K, faces on $\partial\Omega_N \cap \partial K$, and faces on $\partial\Omega_D \cap \partial K$, respectively.

Theorem 1. *Let $\{u^n\}_{1 \leq n \leq \bar{N}}$ and $\{u_h^n\}_{1 \leq n \leq \bar{N}}$ be the semi-discrete solution given by (2) and be the approximate solution given by (3), respectively. Let $1 \leq N \leq \bar{N}$. Then the error $e^n = u^n - u_h^n$, $n = 1, \ldots, N$ satisfies*

$$\| e^N \|_\Omega^2 + \sum_{n=1}^N \tau_n \sum_{K \in \widetilde{\mathscr{T}}_h^n} \| \nabla e^n \|_K^2 \leq \| e^0 \|_\Omega^2 + \sum_{n=1}^N C \left(\tau_n \sum_{K \in \widetilde{\mathscr{T}}_h^n} (\eta_{K,1}^n)^2 + \sum_{K \in \widetilde{\mathscr{T}}_h^n} (\eta_{K,2}^n)^2 \right), \tag{9}$$

where a constant C is independent of the mesh parameter and the time step.

Sketch of the proof. As the whole proof is quite long, we will provide only the main points of it. First, we split the gradient of the error using (4) as follows:

$$\begin{aligned} \tau_n \sum_{K \in \widetilde{\mathscr{T}}_h^n} \| \nabla e^n \|_K^2 = {} & \tau_n \sum_{K \in \widetilde{\mathscr{T}}_h^n} \int_K \nabla e^n \cdot \nabla \phi^n \, \mathrm{d}\mathbf{x} \qquad (=: \psi_1) \\ & + \tau_n \sum_{K \in \widetilde{\mathscr{T}}_h^n} \int_K \nabla e^n \operatorname{curl} \chi^n \, \mathrm{d}\mathbf{x}. \qquad (=: \psi_2) \end{aligned}$$

Setting $\phi := \phi^n$ in (6) multiplied by τ_n and adding a τ_n-multiple of the difference of the right-hand and the left-hand sides of (5) with $v_h := I_h^{n,D}\phi^n$ yield

$$\begin{aligned}\psi_1 =&\tau_n \int_\Omega (f^n - \frac{u^n - u^{n-1}}{\tau_n})\phi^n \,\mathrm{d}\mathbf{x} - \tau_n \sum_{K\in\widetilde{\mathscr{T}}_h^n} \int_{\partial K} \nabla u_h^n \cdot \mathbf{n}\phi^n \,\mathrm{d}S \\ &+ \tau_n \sum_{K\in\widetilde{\mathscr{T}}_h^n} \int_K \Delta u_h^n \phi^n \,\mathrm{d}\mathbf{x} + \tau_n \int_{\partial\Omega_N} g_N^n \phi^n \,\mathrm{d}S - \tau_n \sum_{K\in\widetilde{\mathscr{T}}_h^n} \int_K \nabla e^n \cdot \nabla I_h^{n,D}\phi^n \,\mathrm{d}\mathbf{x} \\ &+ \tau_n \int_\Omega \frac{e^{n-1} - e^n}{\tau_n} I_h^{n,D}\phi^n \,\mathrm{d}\mathbf{x} + \tau_n\theta \sum_{\Gamma\in\widetilde{\mathscr{F}}_h^{n,\mathrm{I}}} \int_\Gamma \langle \nabla I_h^{n,D}\phi^n \cdot \mathbf{n}\rangle [u_h^n] \,\mathrm{d}S.\end{aligned}$$

By expressing term $-\tau_n \sum_{K\in\widetilde{\mathscr{T}}_h^n} \int_K \nabla e^n \cdot \nabla I_h^{n,D}\phi^n \,\mathrm{d}\mathbf{x}$ according to (6), adding and subtracting term $\tau_n \sum_{K\in\widetilde{\mathscr{T}}_h^n} \int_K (f^n - \frac{u_h^n - u_h^{n-1}}{\tau_n})\phi^n \,\mathrm{d}\mathbf{x}$, and reordering the terms, we have

$$\begin{aligned}\psi_1 =&\tau_n \sum_{K\in\widetilde{\mathscr{T}}_h^n} \int_K \mathrm{R}_K^n(\phi^n - I_h^{n,D}\phi^n) \,\mathrm{d}\mathbf{x} \\ &- \sum_{K\in\widetilde{\mathscr{T}}_h^n} \int_K (e^n - e^{n-1})\phi^n \,\mathrm{d}\mathbf{x} - \tau_n \sum_{K\in\widetilde{\mathscr{T}}_h^n} \int_{\partial K} \nabla u_h^n \cdot \mathbf{n}(\phi^n - I_h^{n,D}\phi^n) \,\mathrm{d}S \\ &+ \tau_n \int_{\partial\Omega_N} g_N^n (\phi^n - I_h^{n,D}\phi^n) \,\mathrm{d}S + \tau_n\theta \sum_{\Gamma\in\widetilde{\mathscr{F}}_h^{n,\mathrm{I}}} \int_\Gamma \langle \nabla I_h^{n,D}\phi^n \cdot \mathbf{n}\rangle [u_h^n] \,\mathrm{d}S.\end{aligned} \tag{10}$$

By adding and subtracting suitable terms in (10), estimating all terms in ψ_1 and ψ_2 using approximation properties of $I_h^{n,D}$, trace inequalities, inverse inequality, and well known inequalities such as Hölder's, Young's, etc. together with some auxiliary estimates, we finally come to the assertion of Theorem 1. □

Theorem 2. *Let $\{u^n\}_{1\le n\le \bar{N}}$ and $\{u_h^n\}_{1\le n\le \bar{N}}$ be the semi-discrete solution given by (2) and be the approximate solution given by (3), respectively. Then*

$$h_K \|\mathrm{R}_K^n\|_K \le c\left(h_K\tau_n^{-1}\|e^n - e^{n-1}\|_K + |e^n|_{1,K}\right),\ K \in \widetilde{\mathscr{T}}_h^n$$

$$h_\Gamma^{1/2}\|[\mathbf{n}\cdot\nabla u_h^n]\|_\Gamma \le c(|e^n|_{1,K_\Gamma^\mathrm{L}\cup K_\Gamma^\mathrm{R}} + h_\Gamma\tau_n^{-1}\|e^n - e^{n-1}\|_{K_\Gamma^\mathrm{L}\cup K_\Gamma^\mathrm{R}}),\ \Gamma \in \widetilde{\mathscr{F}}_h^{n,\mathrm{I}}$$

$$h_\Gamma^{1/2}\|g_N^n - \nabla u_h^n\cdot\mathbf{n}\|_\Gamma \le c(|e^n|_{1,K_\Gamma^\mathrm{L}} + h_\Gamma\tau_n^{-1}\|e^n - e^{n-1}\|_{K_\Gamma^\mathrm{L}}),\ \Gamma \in \widetilde{\mathscr{F}}_h^{n,\mathrm{N}}$$

$$C_W\left(J(u_h^n)_{-\frac{1}{2},\widetilde{\mathscr{F}}_h^n}^{u_D^n}\right)^2 \le c\left(\sum_{K\in\widetilde{\mathscr{T}}_h^n}\left(h_K\tau_n^{-1}\|e^n - e^{n-1}\|_K + |e^n|_{1,K}\right)^2\right)$$

where C_W is a constant involved in the penalty parameter σ, K_Γ^{L} and K_Γ^{R} are such that $\Gamma \subset \overline{K_\Gamma^{\mathrm{L}}} \cap \overline{K_\Gamma^{\mathrm{R}}}$, $J(u_h^n)^{u_D^n}_{-\frac{1}{2},\widetilde{\mathscr{F}}_h^n} := \left(\sum_{\Gamma\in\widetilde{\mathscr{F}}_h^{n,\mathrm{I}}} h_\Gamma^{-1} \|[u_h^n]\|_\Gamma^2 + \sum_{\Gamma\in\widetilde{\mathscr{F}}_h^{n,\mathrm{D}}} h_\Gamma^{-1} \|u_h^n - u_D^n\|_\Gamma^2 \right)^{1/2}$, and a constant c is independent of the mesh parameter and the time step.

Sketch of the proof. Due to lack of the space, we will only sketch the proof. As the principle of estimation of the residuum and the normal jumps terms is the same, we will show the derivation only for the residuum. Summing (6) and (7) and rewritten it in more convenient form leads to

$$\sum_{K\in\widetilde{\mathscr{T}}_h^n} \int_K \nabla e^n (\nabla\phi + \operatorname{curl}\chi) dx + \int_\Omega \frac{e^n - e^{n-1}}{\tau_n}\phi dx = -\sum_{\Gamma\in\widetilde{\mathscr{F}}_h^{n,\mathrm{I}}} \int_\Gamma [\nabla u_h^n \cdot \mathbf{n}]\phi dS$$

$$+ \sum_{K\in\widetilde{\mathscr{T}}_h^n} \int_K \mathrm{R}_K^n \phi dx + \int_{\partial\Omega_N} (g_N^n - \nabla u_h^n \cdot \mathbf{n})\phi dS + \sum_{K\in\widetilde{\mathscr{T}}_h^n} \int_{\partial K\setminus\partial\Omega_N} e^n \operatorname{curl}\chi \cdot \mathbf{n} dS, \quad (11)$$

for $\phi \in H_D^1(\Omega)$ and $\underset{\sim}{\chi} \in (H^1(\Omega))^k$ ($k = 1$ for $d = 2$ and $k = 3$ for $d = 3$). Let fix an arbitrary $K \in \widetilde{\mathscr{T}}_h^n$. Setting $\phi|_K := b_K \mathrm{R}_K^n$, $\phi := 0$ outside of K and $\chi := 0$ in (11), where b_K is a standard interior bubble function supported on the element K, yields

$$\int_K \mathrm{R}_K^n b_K \mathrm{R}_K^n dx = \int_K \frac{e^n - e^{n-1}}{\tau_n} b_K \mathrm{R}_K^n dx + \int_K \nabla e^n \cdot \nabla(b_K \mathrm{R}_K^n) dx. \quad (12)$$

Further, there exists a constant $c > 0$ such that the inequality $\int_K (\mathrm{R}_K^n)^2 dx \le c \int_K \mathrm{R}_K^n b_K \mathrm{R}_K^n dx$ holds, because $(\int_K (\cdot)^2 b_K dx)^{1/2}$ is a norm on $L^2(K)$ ($b_K > 0$ on the interior of K), equivalent to the L^2 norm on $P^p(K)$.

Now, applying the Cauchy-Schwarz inequality together with the inverse inequality $|b_K \mathrm{R}_K^n|_{1,K} \le h_K^{-1} \|\mathrm{R}_K^n\|_K$ in (12), we obtain

$$\|\mathrm{R}_K^n\|_K^2 \le c \left(\tau_n^{-1} \left\| e^n - e^{n-1} \right\|_K + |e^n|_{1,K} h_K^{-1} \right) \|\mathrm{R}_K^n\|_K.$$

According to (8), it remains to estimate $\sum_{\Gamma\in\widetilde{\mathscr{F}}_K^{n,\mathrm{I}}} h_\Gamma^{-1/2} \|[u_h^n]\|_\Gamma + \sum_{\Gamma\in\widetilde{\mathscr{F}}_K^{n,\mathrm{D}}} h_\Gamma^{-1/2} \|u_D^n - u_h^n\|_\Gamma$, which has been done in [4]. □

4 Numerical Example

In this section, we present numerical experiments illustrating the a posteriori error estimates of this paper. We consider the problem (1) where $T = 1$, $\Omega = (0,1) \times (0,1)$, $\partial\Omega_N = \emptyset$, and the initial and boundary conditions are chosen in such a way

Table 1 The computed errors, error estimators, and effectivity indices

P_k	h_m	τ_m	$\|e_h\|_Y$	η_1	η_2	η_{IC}	η_{tot}	i_{eff}
1	1.25E−01	1.00E−02	1.22E+00	1.43E+01	3.01E−01	4.31E−03	1.43E+01	11.7360
1	6.25E−02	5.00E−03	6.10E−01	7.39E+00	1.00E−01	1.08E−03	7.39E+00	12.1190
	(EOC)		(1.00)	(0.95)	(1.59)	(2.00)	(0.95)	
1	3.12E−02	2.50E−03	3.05E−01	3.76E+00	3.40E−02	2.70E−04	3.76E+00	12.3178
	(EOC)		(1.00)	(0.97)	(1.56)	(2.00)	(0.97)	
1	1.56E−02	1.25E−03	1.53E−01	1.90E+00	1.18E−02	6.75E−05	1.90E+00	12.4193
	(EOC)		(1.00)	(0.99)	(1.53)	(2.00)	(0.99)	
2	1.25E−01	1.00E−02	2.02E−01	4.53E−01	9.96E−03	7.89E−05	4.53E−01	2.2479
2	6.25E−02	2.50E−03	5.04E−02	1.15E−01	2.46E−03	9.88E−06	1.15E−01	2.2898
	(EOC)		(2.00)	(1.97)	(2.02)	(3.00)	(1.97)	
2	3.12E−02	6.25E−04	1.26E−02	2.92E−02	6.10E−04	1.24E−06	2.92E−02	2.3152
	(EOC)		(2.00)	(1.98)	(2.01)	(3.00)	(1.98)	
2	1.56E−02	1.56E−04	3.15E−03	7.34E−03	1.52E−04	1.54E−07	7.35E−03	2.3338
	(EOC)		(2.00)	(1.99)	(2.00)	(3.00)	(1.99)	
3	1.25E−01	1.00E−02	1.99E−01	1.21E−02	1.66E−04	1.11E−06	1.21E−02	0.0609
3	6.25E−02	1.25E−03	2.49E−02	1.46E−03	2.43E−05	6.95E−08	1.49E−03	0.0598
	(EOC)		(3.00)	(3.04)	(2.77)	(4.00)	(3.03)	
3	3.12E−02	1.56E−04	3.11E−03	1.84E−04	4.08E−06	4.34E−09	1.96E−04	0.0630
	(EOC)		(3.00)	(2.99)	(2.58)	(4.00)	(2.93)	

that the exact solution is $u(x_1, x_2, t) = \exp[x_1 + x_2 + 2t]$. We simply observe that the right-hand side f of (1) vanishes. We performed a set of numerical experiments with the aid of the DGM (3) for $p = 1, 2, 3$ polynomial approximations.

We consider a uniform space-time discretizations characterized by the space and time steps h_m and τ_m, $m = 1, \ldots, 4$, respectively. We choose $\{h_1, \tau_1\} = (1/8, 1/100)$ and then set $h_{m+1} = h_m/2$, $\tau_{m+1} = \tau_m/2^p$ for $m = 1, 2, 3$. The space grids are triangulations with right-angled triangles resulting from diagonal cuttings of squares with edges of the length $h_l = h_K/\sqrt{2}$. We evaluate the experimental order of convergence $\mathrm{EOC} := \frac{\log(E_m/E_{m-1})}{\log(h_m/h_{m-1})}$, $m = 2, 3, 4$, where E_m is either an error, or an error estimator on the space-time discretization $\{h_m, \tau_m\}$. Table 1 shows the values from (9), namely

$$\|e_h\|_Y := \|e^N\|_\Omega^2 + \sum_{n=1}^{N} \tau_n \sum_{K \in \widetilde{\mathscr{T}}_h^n} \|\nabla e^n\|_K^2, \; \eta_1 := \sum_{n=1}^{N} \tau_n \sum_{K \in \widetilde{\mathscr{T}}_h^n} (\eta_{K,1}^n)^2,$$

$$\eta_2 := \sum_{n=1}^{N} \sum_{K \in \widetilde{\mathscr{T}}_h^n} (\eta_{K,2}^n)^2, \; \eta_{\mathrm{IC}} := \|e^0\|_\Omega^2, \; \eta_{\mathrm{tot}} := \eta_1 + \eta_2 + \eta_{\mathrm{IC}}, \; i_{\mathrm{eff}} := \eta_{\mathrm{tot}}/\|e_h\|_Y.$$

The value i_{eff} corresponds to the *effectivity index*. However, since our estimate (9) contains an undetermined constant C, this value may be lower than one. As Table 1 shows, AEE is independent of the discretization parameters h and τ but depends on the degree of polynomial approximation p. Similar observation has been made in [5], where i_{eff} tends to a fixed value for $h, \tau \to 0$, but it differs for different p.

References

1. Arnold, D.N., Brezzi, F., Cockburn, B., Marini, L.D.: Unified analysis of discontinuous Galerkin methods for elliptic problems. SIAM J. Numer. Anal. **39**(5), 1749–1779 (2001/02)
2. Dari, E., Durán, R., Padra, C., Vampa, V.: A posteriori error estimators for nonconforming finite element methods. RAIRO Modél. Math. Anal. Numér. **30**(4), 385–400 (1996)
3. Dolejší, V., Šebestová, I.: A posteriori error estimates of the discontinuous Galerkin method for the heat conduction equation (2011). Accepted to AUC
4. Ern, A., Vohralík, M.: A posteriori error estimation based on potential and flux reconstruction for the heat equation. SIAM J. Numer. Anal. **48**(1), 198–223 (2010)
5. Georgoulis, E.H., Lakkis, O., Virtanen, J.M.: A posteriori error control for discontinuous Galerkin methods for parabolic problems. SIAM J. Numer. Anal. **49**(2), 427–458 (2011)
6. Šebestová I.: A posteriori error estimates of the discontinuous Galerkin method for convection-diffusion equations. Master thesis, Charles University in Prague (2009). URL atrey.karlin.mff.cuni.cz/~dolejsi/Ivana/DP.pdf
7. Karakashian, O.A., Pascal, F.: A posteriori error estimates for a discontinuous Galerkin approximation of second-order elliptic problems. SIAM J. Numer. Anal. **41**(6), 2374–2399 (2003)
8. Nicaise, S., Soualem, N.: A posteriori error estimates for a nonconforming finite element discretization of the heat equation. M2AN Math. Model. Numer. Anal. **39**(2), 319–348 (2005)
9. Verfürth, R.: A review of a posteriori error estimation and adaptive mesh-refinement techniques. Teubner-Wiley, Stuttgart (1996)
10. Verfürth, R.: A posteriori error estimates for finite element discretizations of the heat equation. Calcolo **40**(3), 195–212 (2003)

Distributed Optimal Control of Diffusion-Convection-Reaction Equations Using Discontinuous Galerkin Methods

H. Yücel, M. Heinkenschloss, and B. Karasözen

Abstract We discuss the symmetric interior penalty Galerkin (SIPG) method, the nonsymmetric interior penalty Galerkin (NIPG) method, and the incomplete interior penalty Galerkin (IIPG) method for the discretization of optimal control problems governed by linear diffusion-convection-reaction equations. For the SIPG discretization the *discretize-then-optimize (DO)* and the *optimize-then-discretize (OD)* approach lead to the same discrete systems and in both approaches the observed L^2 convergence for states and controls is $O(h^{k+1})$, where k is the degree of polynomials used. The situation is different for NIPG and IIPG, where the the DO and the OD approach lead to different discrete systems. For example, when standard penalization is used, the L^2 error in the controls is only $O(h)$ independent of k. However, if superpenalization is used, the lack of adjoint consistency is reduced and the observed convergence for NIPG and IIPG is essentially equal to that of the SIPG method in the DO and OD approach.

1 Introduction

We study converge rates of three discontinuous Galerkin (DG) discretizations of linear-quadratic optimal control problems governed by diffusion-convection-reaction partial differential equations (PDEs) with distributed controls using both the *discretize-then-optimize (DO)* and the *optimize-then-discretize (OD)* approach.

H. Yücel (✉) · B. Karasözen
Department of Mathematics and Institute of Applied Mathematics, Middle East Technical University, 06800, Ankara, Turkey
e-mail: hayucel@metu.edu.tr; bulent@metu.edu.tr

M. Heinkenschloss
Department of Computational and Applied Mathematics, Rice University, MS-134, 6100 Main Street, Houston, 77005-1892, TX, USA
e-mail: heinken@rice.edu

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_42,

In the DO approach one first discretizes the optimal control problem (the state PDE and the objective function) and then one forms the linear system of optimality conditions. In the OD approach one first forms the system of optimality conditions, which consists of the state PDE, the adjoint PDE, and, for our problems, an algebraic equation that links the control and the adjoint variable. Afterwards one discretizes this optimality PDE system to derive a linear system. For optimal control problems governed by diffusion-convection-reaction PDEs, it is known that for some discretization schemes the two approaches lead to the same linear systems (this is, e.g., the case for local projection based stabilization [3] and edge stabilization [11] of continuous Galerkin discretizations), whereas for other discretization schemes the two approaches lead to different linear systems (this is, e.g., the case for the streamline upwind Petrov Galerkin (SUPG) method [5, 8]). In the latter case it is important to understand what the differences are and how they impact the discretization error. The goal of this paper is study this issue for three DG methods, the symmetric interior penalty Galerkin (SIPG) method, the nonsymmetric interior penalty Galerkin (NIPG) method, and the incomplete interior penalty Galerkin (IIPG) method.

We will show that when the SIPG method is used, both approaches lead to the same result, whereas different systems arise when the NIPG method or the IIPG method is used. This matches the results in [9]. Consequently, when the NIPG or the IIPG method are used, there can be significant differences in the convergence of the discretized solution as the mesh is refined depending on whether DO or the OD approach is used. The OD approach gives better convergence rates. However, when we use superpenalization, then we observe the same convergence behavior for the DO and the OD approaches. This is important since the DO approach has an important advantage: In the linear-quadratic case it leads to symmetric optimality systems. In the nonlinear, non-quadratic case, the DO leads to consistent gradients, which makes the application of gradient based optimization methods easier. (Gradient approximation computed by the OD approach are only asymptotically consistent.) Our results indicate that if the NIPG or the IIPG method is used, the DO approach can be used without suffering a deterioration in the convergence as the mesh is refined, provided superpenalization is applied.

2 The Optimal Control Problem

Let Ω be a bounded open, convex domain in $\mathbb{R}^2$ with boundary $\Gamma = \partial\Omega$, let $f, y_d \in L^2(\Omega)$, $g_D \in H^{3/2}(\Gamma)$, $\mathbf{c} \in \left(W^{1,\infty}(\Omega)\right)^2$, and $r \in L^\infty(\Omega)$, be given functions with $r - \frac{1}{2}\nabla \cdot \mathbf{c} \geq 0$ a.e. in Ω, and let $\epsilon, \omega > 0$ be given scalars. We consider the linear quadratic optimal control problem

$$\text{minimize } J(y,u) := \frac{1}{2}\int_\Omega (y(x) - y_d(x))^2 dx + \frac{\omega}{2}\int_\Omega u(x)^2 dx \qquad (1)$$

subject to

$$-\epsilon\Delta y(x) + \mathbf{c}(x)\cdot\nabla y(x) + r(x)y(x) = f(x) + u(x), \qquad x \in \Omega, \quad (2a)$$

$$y(x) = g_D(x), \qquad x \in \Gamma, \quad (2b)$$

We refer to u as the control, to y as the state and to (2) as the state equation.

The state $y \in H^1(\Omega)$ and the control $u \in L^2(\Omega)$ solve the optimal control problem (1) and (2) if and only if there exists an adjoint $p \in H^1(\Omega)$ such that y, u, p satisfy the state equation (2), the adjoint equation

$$-\epsilon\Delta p(x) - \mathbf{c}(x)\cdot\nabla p(x) + (r(x) - \nabla\cdot\mathbf{c}(x))p(x) = -(y(x) - y_d(x)), \quad x \in \Omega, \quad (3a)$$

$$p(x) = 0, \qquad x \in \Gamma, \quad (3b)$$

and

$$\omega u(x) - p(x) = 0, \qquad x \in \Omega. \quad (4)$$

Regularity results for the solution y, u, p of (2)–(4) and dependence of the H^1, H^2 norms of y, u, p on ϵ are given in [6].

If the reduced formulation of (1) and (2) is used, i.e., if the solution y of the state equation (2) is viewed as an implicit function of the control u and $\widehat{J}(u) = J(y(u), u)$ is minimized, then $\nabla\widehat{J}(u) = \omega u - p$, where p solves (3) with $y = y(u)$.

DG methods have many advantages over continuous finite element methods for the numerical solution of convection-diffusion-reaction equations. Although there are several papers which analyze DG methods applied to (2) with fixed u, there are few papers dealing with the application of DG methods to optimal control problems governed by diffusion-convection-reaction equation [8, 12].

3 Discontinuous Galerkin Discretization

For the statement of the DG methods we use the notation in Chaps. 2 and 4 of [10] and for the spaces of the state, adjoint, control variables and test functions we use $U_h = V_h = Y_h = \{y \in L^2(\Omega) : y|_E \in \mathbb{P}_k(E),\ \forall E \in \mathscr{E}_h\}$ [8].

3.1 Discretization of State Equation

In this section we review our DG discretizations of the state equation (2) for a fixed control u. The diffusion part is discretized using SIPG, NIPG, or IIPG and the convection part is discretized by a upwind method [2, 7, 8]. This leads to

$$
\begin{aligned}
a_h^s(y_h,\upsilon_h) = & \sum_{E\in\mathscr{E}_h} (\epsilon\nabla y_h, \nabla\upsilon_h)_E \\
& + \kappa \sum_{e\in\Gamma_h} (\{\epsilon\nabla\upsilon_h\cdot n_e\}, [y_h])_e - \sum_{e\in\Gamma_h} (\{\epsilon\nabla y_h\cdot n_e\}, [\upsilon_h])_e \\
& + \sum_{e\in\Gamma_h} \frac{\sigma\epsilon}{h_e^{\beta_0}} ([y_h],[\upsilon_h])_e + \sum_{E\in\mathscr{E}_h} (\mathbf{c}\cdot\nabla y_h + r y_h, \upsilon_h)_E \\
& + \sum_{e\in\Gamma_h^0} (y_h^+ - y_h^-, |n\cdot\mathbf{c}|\upsilon_h^+)_e + \sum_{e\in\Gamma_h^-} (y_h^+, \upsilon_h^+|n\cdot\mathbf{c}|)_e, \qquad (5a)
\end{aligned}
$$

$$
b_h(u_h,\upsilon_h) = -\sum_{E\in\mathscr{E}_h} (u_h,\upsilon_h)_E, \qquad (5b)
$$

$$
\begin{aligned}
l_h^s(\upsilon_h) = & \sum_{E\in\mathscr{E}_h} (f,\upsilon_h)_E + \sum_{e\in\Gamma_h^\partial} \left(\frac{\sigma\epsilon}{h_e^{\beta_0}} (g_D,[\upsilon_h])_e + \kappa(\epsilon g_D, \nabla\upsilon_h\cdot n)_e \right) \\
& + \sum_{e\in\Gamma_h^-} (g_D, \upsilon_h^+|n\cdot\mathbf{c}|)_e. \qquad (5c)
\end{aligned}
$$

Depending on the choices of κ one obtains several variants of DG methods: SIPG ($\kappa = -1$), IIPG ($\kappa = 0$), NIPG ($\kappa = 1$). The nonnegative real parameter σ is called the penalty parameter. The choice of penalty parameter is crucial for the convergence of discontinuous Galerkin methods (see e.g. [10]). For NIPG $\sigma = 1$ and for SIPG and IIPG σ must be sufficiently large [10, Sect. 2.7.1]. Furthermore, $\beta_0 = 1$ in standard penalization.

In (5) the superscript s is used to indicate that the DG methods are applied to the state equation. The discontinuous Galerkin discretization of the state equation (2) for a fixed control $u_h \in U_h$ is given by the following variational form: Find $y_h \in Y_h$ such that

$$
a_h^s(y_h,\upsilon_h) + b_h(u_h,\upsilon_h) = l_h^s(\upsilon_h), \qquad \forall \upsilon_h \in V_h. \qquad (6)
$$

It is shown in [10, Theorem 2.14, p. 48] that when the Dirichlet data g_D is a continuous piecewise polynomial of degree k and the exact solution to (2) belongs to $H^s(\mathscr{E}_h)$ for $s > 3/2$, then the following error estimate is obtained:

$$
\|y - y_h\|_{L^2(\Omega)} \le C h^{\min(k+1,s)} \|\|y\|\|_{H^s(\mathscr{E}_h)}, \qquad (7)
$$

where C independent of h. Recall that k is the degree of polynomials. This estimate is valid for SIPG unconditionally and for the NIPG and IIPG methods provided the superpenalization $\beta_0 \ge 3$ is applied [1, 4, 10]. In particular, when superpenalization is used the NIPG and IIPG methods produce same convergence rates as SIPG.

3.2 Discretize-Then-Optimize Approach

When we discretize the optimal control problem using the DG methods discussed previously, we arrive at

$$\text{minimize} \quad J(y_h, u_h) := \frac{1}{2} \sum_{E \in \mathscr{E}_h} \|y_h - y_d\|_E^2 + \frac{\omega}{2} \sum_{E \in \mathscr{E}_h} \|u_h\|_E^2, \tag{8a}$$

$$\text{subject to} \quad a_h^s(y_h, \upsilon_h) + b_h(u_h, \upsilon_h) = l_h^s(\upsilon_h), \qquad \forall \upsilon_h \in V_h,$$

$$(y_h, u_h) \in Y_h \times U_h. \tag{8b}$$

The necessary and sufficient optimality conditions for (8) are

$$a_h^s(\psi_h, p_h) = -(y_h - y_d, \psi_h), \quad \forall \psi_h \in V_h, \tag{9a}$$

$$b_h(w_h, p_h) + \omega(u_h, w_h) = 0, \qquad \forall w_h \in U_h, \tag{9b}$$

$$a_h^s(y_h, \upsilon_h) + b_h(u_h, \upsilon_h) = l_h^s(\upsilon_h), \qquad \forall \upsilon_h \in V_h. \tag{9c}$$

3.3 Optimize-Then-Discretize Approach

In this approach we discretize the optimality system (2)–(4) directly. The DG discretization of the state equation (2) is the same as in (9). Since the adjoint equation (3) is also an convection-diffusion-reaction equation, but with negative convection term $-\mathbf{c} \cdot \nabla p$ we use the apply the same DG method that is used for the state. This leads to the discretized adjoint equations

$$a_h^a(p_h, \psi_h) = -(y_h - y_d, \psi_h) \qquad \forall \psi_h \in V_h, \tag{10}$$

where

$$\begin{aligned} a_h^a(p_h, \psi_h) = & \sum_{E \in \mathscr{E}_h} (\epsilon \nabla p_h, \nabla \psi_h)_E \\ & + \kappa \sum_{e \in \Gamma_h} (\{\epsilon \nabla \psi_h \cdot n_e\}, [p_h])_e - \sum_{e \in \Gamma_h} (\{\epsilon \nabla p_h \cdot n_e\}, [\psi_h])_e \\ & + \sum_{e \in \Gamma_h} \frac{\sigma \epsilon}{h_e^{\beta_0}} ([p_h], [\psi_h])_e + \sum_{E \in \mathscr{E}_h} (-\mathbf{c} \cdot \nabla p_h + (r - \nabla \cdot \mathbf{c}) p_h, \psi_h)_E \\ & + \sum_{e \in \Gamma_h^0} (p_h^+ - p_h^-, |n \cdot \mathbf{c}| \psi_h^+)_e + \sum_{e \in \Gamma_h^+} (p_h^+, \psi_h^+ |n \cdot \mathbf{c}|)_e. \end{aligned}$$

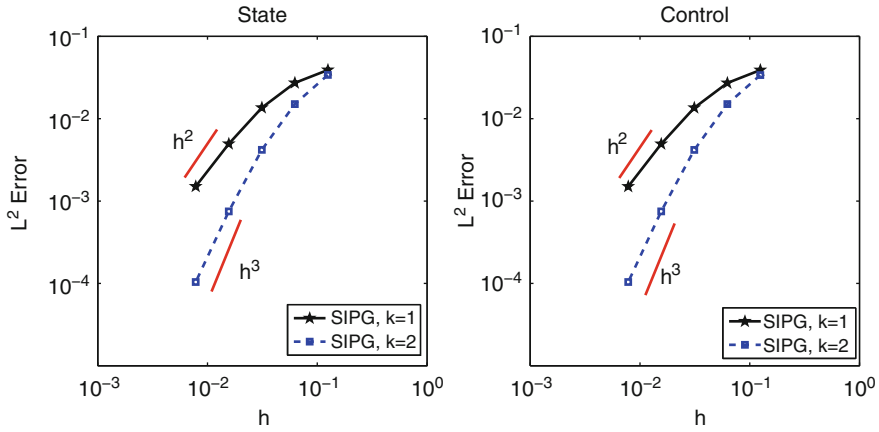


Fig. 1 The L_2 state and control errors for the SIPG discretization

The superscript a indicates that the DG method is applied to the adjoint equation. Since the adjoint variable is equal to zero on the boundary Γ, there is no additional term at the right-side. Finally, the DG discretization of (4) is the same as in (9).

If $a_h^a(p_h, \psi_h) = a_h^s(\psi_h, p_h)$, then the optimality system (9) of the DG discretized optimal control problem (8) is equivalent to the DG discretization (10), (9b) and (9c) of the optimality system (2)–(4). For SIPG, we have $a_h^s(\upsilon_h, p_h) = a_h^a(p_h, \upsilon_h)$ for all $p_h, \upsilon_h \in V_h$, whereas for NIPG and IIPG $a_h^s(\upsilon_h, p_h) \neq a_h^a(p_h, \upsilon_h), \forall \upsilon_h \in V_h$.

4 Numerical Results

We use piecewise linear ($k = 1$) and piecewise quadratic ($k = 2$) polynomials. The penalty parameter is $\sigma = 1$ for all edges for NIPG. For SIPG and IIPG we set $\sigma = 3k(k+1)$ for interior edges and $\sigma = 6k(k+1)$ on boundary edges.

If standard penalization $\beta_0 = 1$ is used in NIPG and IIPG, we refer to these methods as NIPG1 and IIPG1, respectively. If superpenalization $\beta_0 = 3$ is used, we refer to these methods as NIPG3 and IIPG3, respectively.

Consider the following distributed optimal control problem from [5]. Let $\mathbf{c} = (\cos(\theta), \sin(\theta))^T$, $\theta = 45°$, $r = 0$, $\epsilon = 10^{-2}$ and $\omega = 1$ in $\Omega = (0,1)^2$. The functions f, y_d and g_D are chosen such that the exact state and adjoint are $y_{ex}(x_1, x_2) = \eta(x_1)\eta(x_2)$ and $p_{ex}(x_1, x_2) = \xi(x_1)\xi(x_2)$ with

$$\eta(z) = z - \frac{\exp((z-1)/\epsilon) - \exp(-1/\epsilon)}{1 - \exp(-1/\epsilon)}, \quad \xi(z) = 1 - z - \frac{\exp(-z/\epsilon) - \exp(-1/\epsilon)}{1 - \exp(-1/\epsilon)}.$$

If the SIPG method is used, the DO and the OD approach are identical. The L^2 convergence results are shown in Fig. 1. The observed convergence rate is $O(h^{k+1})$, which is expected from the convergence of SIPG for a single PDE (see, e.g., (7) and [10]).

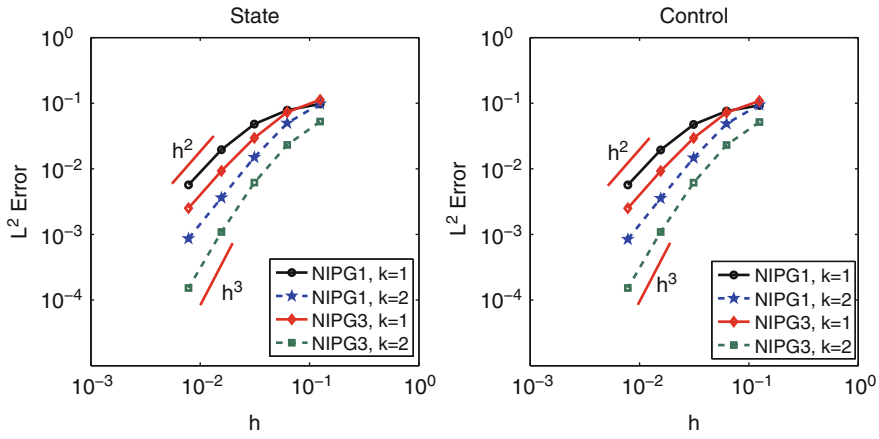


Fig. 2 L_2 state and control errors for the NIPG1 and NIPG3 discretization using the OD approach

The situation is different for NIPG and IIPG. Because of page limitations we only show NIPG results; those for IIPG are similar. For NIPG the DO and the OD approaches lead to different discrete optimality systems. We observe that for NIPG1, the OD approach exhibits better convergence. In particular, in the OD approach using NIPG1 the observed convergence rates for the controls are $O(h^k)$ (see Fig. 2). In contrast, if the DO approach is used, the observed convergence rates for the controls are $O(h)$ for both $k = 1$ and $k = 2$ (see Fig. 3). However, if superpenalization is used, the convergence behavior of NIPG3 discretization is essentially the same for the DO and the OD approach. In both cases the observed convergence rates for state and controls is $O(h^{k+1})$. See Figs. 2 and 3. This is not completely surprising, since when a large penalty parameter is used the lack of adjoint consistency is reduced [10].

We also note that there is price to pay for using superpenalization. Superpenalization potentially increases the condition number of the discrete optimality system. For example, for a single elliptic PDE the condition number of the discrete system is of order $O(h^{-2})$ for NIPG1 and of order $O(h^{-4})$ for NIPG3 [4] and [10, p. 59]. Large penalty parameters also decrease the jumps across element interfaces, which can affect the stability of the NIPG and IIPG methods with superpenalization for truly convection dominated problems. Since this problem class is the target, it is ultimately favorable to use a scheme like SIPG for which optimize-then-discretize and discretize-then-optimize commute, rather than to find a penalty parameter for NIPG and IIPG that reduces the effect of adjoint inconsistency while maintaining the stability properties of these methods for convection dominated problems.

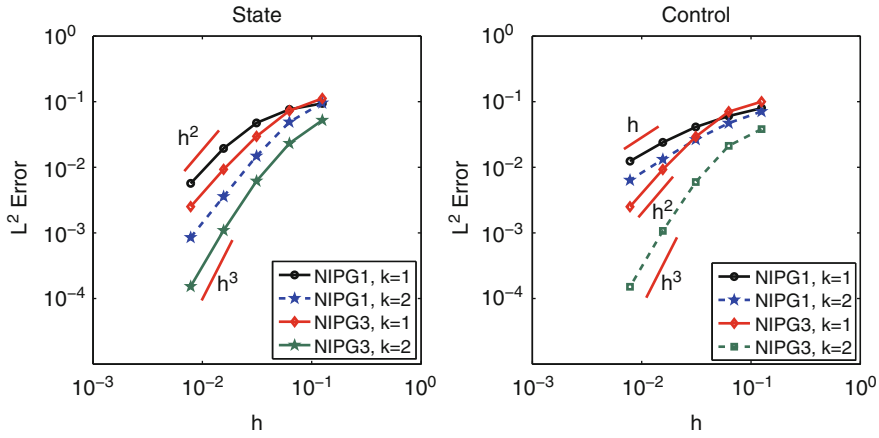


Fig. 3 L_2 state and control errors for the NIPG1 and NIPG3 discretization using the DO approach

Acknowledgements HY has been supported by the 2214-International Doctoral Research Fellowship Program TÜBITAK during his studies in the Department of Computational and Applied Mathematics, Rice University, Houston. The work of MH was supported in part by NSF DMS-0915238. BK was supported through a Fulbright Scholarship.

References

1. Arnold D., Brezzi F., Cockborn B., Marini L.: Unified analysis of discontinuous Galerkin methods for elliptic problems. SIAM J. Numer. Anal. 39, 1749–1779 (2002).
2. Ayuso B., Marini L. D.: Discontinuous Galerkin methods for advection-diffusion-reaction problems. SIAM J. Numer. Anal. 47, 1391–1420 (2009).
3. Becker R., Vexler B.: Optimal control of the convection-diffusion equation using stabilized finite element methods. Numer. Math. 106, 349–367 (2007).
4. Castillo P.: Performance of Discontinuous Galerkin Methods for Elliptic PDEs. SIAM J. Sci. Comput. 24, 624–547 (2002).
5. Collis S. S., Heinkenschloss M.: Analysis of the streamline upwind/Petrov Galerkin method applied to the solution of optimal control problems. Tech. Rep. TR02-01, Department of Computational and Applied Mathematics, Rice University, (2002).
6. Heinkenschloss M., Leykekhman D.: Local error estimates for SUPG solutions of advection-dominated elliptic linear-quadratic optimal control problems. SIAM J. Numer. Anal. 47, 4607–4638 (2010).
7. Houston P., Schwab C., Süli E.: Discontinuous hp-Finite Element Methods for Advection-Diffusion-Reaction Problems. SIAM J. Numer. Anal. 39, pp. 2133–2163 (electronic) (2002).
8. Leykekhman D., Heinkenschloss M.: Local error analysis of discontinuous Galerkin methods for advection-dominated elliptic linear-quadratic optimal control problems. SIAM J. Numer. Anal. 50, 2012–2038 (2012).
9. Leykekhman D.: Investigation of Commutative Properties of Discontinuous Galerkin Methods in PDE Constrained Optimal Control Problems. J. of Scientific Computing, 1–29 (2012).

10. Rivière B.: Discontinuous Galerkin methods for solving elliptic and parabolic equations. Theory and implementation. SIAM Volume 35 of Frontiers in Applied Mathematics, (2008).
11. Yan, N, Zhou, Z.: A priori and a posteriori error analysis of edge stabilization Galerkin method for the optimal control problems governed by convection-dominated diffusion equation, Journal of Computational and Applied Mathematics 223, 198–217 (2009).
12. Zhou Z., Yan N.: The local discontinuous Galerkin method for optimal control problem governed by convection-diffusion equations. International Journal of Numerical Analysis & Modeling 7, 681–699 (2010).

An Immersed Boundary Method for Drug Release Applied to Drug Eluting Stents Dedicated to Arterial Bifurcations

L. Cattaneo, C. Chiastra, E. Cutrì, F. Migliavacca, S. Morlacchi, and P. Zunino

Abstract We address an immersed boundary method applied to the study of cardiovascular drug eluting stents deployed in coronary bifurcations. The problem involves the interaction of arterial deformations, hemodynamics and controlled drug release. Resorting to an immersed boundary method facilitates the handling of complex stent pattern and simplifies the definition of the mathematical model for drug release.

1 Introduction

The treatment of coronary bifurcation lesions, occurring up to 16 % of coronary stenosis [1, 8] represents a challenge for the interventional cardiologists due to the lower rate of procedural success and the higher rate of restenosis. The advent of drug eluting stents (DES) in stenting intervention for coronary artery diseases has dramatically reduced restenosis and consequently the request for re-intervention [11]. Despite these advantages, the conventional double stenting techniques can lead to stent underexpansion at the side branch ostium, partial coverage at the side branch origin, incomplete stent apposition and an high density of the metal at the carina [3].

L. Cattaneo · E. Cutrì · P. Zunino (✉)
MOX – Department of Mathematics, Politecnico di Milano, p.zza Leonardo da Vinci 32, 20133, Milano, Italy
e-mail: paolo.zunino@polimi.it

C. Chiastra · S. Morlacchi
LaBS – Department of Structural Engineering, Bioengineering Department, Politecnico di Milano, p.zza Leonardo da Vinci 32, 20133, Milano, Italy

F. Migliavacca
LaBS – Department of Structural Engineering, Politecnico di Milano, p.zza Leonardo da Vinci 32, 20133, Milano, Italy

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_43,

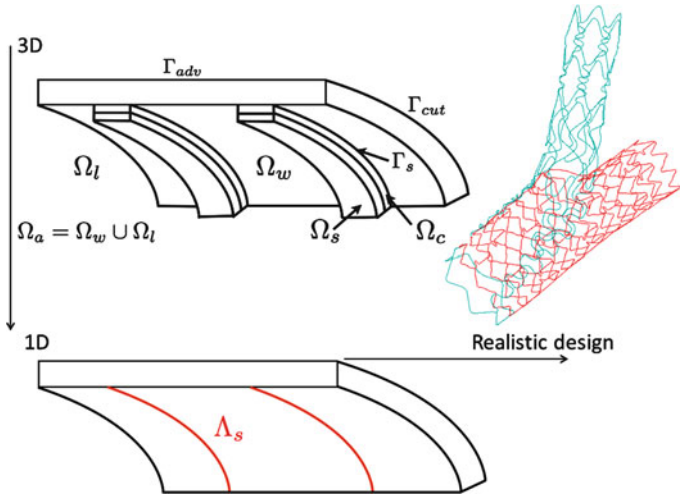


Fig. 1 A sketch of the domain of analysis with labels for subregions and boundaries (*top-left*); the 1D geometrical description of the centerline Λ_s at the basis of the immersed boundary model (*bottom-left*); a realistic stent design applied in the numerical simulations (*right*)

In order to overcome these limitations dedicated devices to arterial bifurcation has been developed.

Computational modeling of these devices is a challenging task, because of their complex geometrical pattern (see Fig. 1 for a realistic example) coupled to the multi-physics nature of the problem, involving the interaction among arterial deformations, hemodynamics and controlled drug release. We remand the interested reader to [7, 14] for an overview. As illustrated in [9], the application of immersed boundary methods is particularly helpful in such context. In our case, the immersed boundary formulation facilitates the handling of the complex stent pattern and allows us to considerably simplify the formulation of the mathematical model for drug release.

2 An Immersed Boundary Model for Drug Release

For the modeling of the artery we consider a computational domain $\Omega_a \subset \mathbb{R}^3$, given by a truncated portion of an artery including both the lumen and the arterial wall, i.e. $\Omega_a = \Omega_w \cup \Omega_l$, where the subscript w denotes the arterial wall and l corresponds to the lumen. The boundary $\partial\Omega_a$ can be split into Γ_{adv}, the interface with the outer wall tissue, called adventitia, and Γ_{cut}, representing the artificial sections where the artery has been truncated from the entire vascular system. For the modeling of the stent we consider a computational domain $\Omega_s \subset \mathbb{R}^3$, we assume that DES beams (also called *struts*) have a circular section and we define as ρ_s its radius; if the real strut features a rectangular section with perimeter P, the equivalent radius is given

by $\rho_s = P/(2\pi)$. For the modeling of drug release we consider a domain $\Omega_c \subset \mathbb{R}^3$ corresponding to the thin substrate surrounding the stent and releasing drug and we denote by $\Gamma_s = \partial\Omega_a \cap \partial\Omega_c$, the interface with the artery.

To avoid resolving the complex 3D geometry of the stent we use an immersed boundary method, which takes into account only the 1D geometrical description of the centerline Λ_s of the stent, as depicted in Fig. 1. The idea of this method is to define an asymptotic problem, applying a suitable rescaling and let the equivalent radius of the stent beams, denoted by ρ_s, vanish in order to replace an immersed interface and the related interface conditions by an equivalent mass source capable to release drug. The advantage of such an approach relies in its efficiency, because it does not need a full description of the stent geometry allowing for a large economy of memory and CPU time, making it possible to reproduce different stent configurations without affecting the computational mesh for approximating fluid dynamics and mass transport.

Regarding the model set up, we denote by f the mass flux per unit area released by the surface Γ_s. From the modeling point of view, if $a(t, \mathbf{x})$ with $\mathbf{x} \in \Gamma_s$ is the drug concentration in the artery, then $f(t, a(t, \mathbf{x}))$ is a pointwise constitutive law for the drug release rate. This means that at the interface between the drug substrate and the artery we should impose that:

$$D\nabla a(t, \mathbf{x}) \cdot \mathbf{n} = f(t, a(t, \mathbf{x})) \qquad \text{on } \Gamma_s,$$

where D is the drug diffusivity in the artery and $\mathbf{n}$ is the outer unit normal vector with respect to Γ_s. The immersed boundary method is able to represent the action of f on Γ_s as an equivalent source term, F, distributed on the entire domain Ω_a. More precisely, $F = F(t, a)$ is a measure defined by

$$\int_{\Omega_a} F(t, a)v = \int_{\Gamma_s} f(t, a)v \qquad \forall v \in C(\Omega), \tag{1}$$

where v plays the role of a test function in the variational setting. Hence, we use the notation $F(t, a) = f(t, a)\delta_{\Gamma_s}$, meaning that F is the Dirac measure concentrated on Γ_s, having (time and concentration dependent) density f on Γ_s. Following along the lines of [4], when $\rho_s \to 0$ we represent the mass flux per unit area by an equivalent mass flux per unit length, distributed on the centerline Λ_s of the stent and exploiting numerical integration we approximate the action of F on v in (1) by means of an integral with respect to the arc length of Γ_s. More precisely, using cylindrical coordinates (s, θ) on Γ_s, we have

$$\int_{\Omega} F(t, a)v = \int_{\Lambda_s} \int_{\gamma(s)} f(t, a(t, s, \rho_s, \theta))v(s, \rho_s, \theta)\rho_s d\theta ds. \tag{2}$$

To approximate the integral of a function g over the arc $\gamma(s)$ we apply the (midpoint) rectangle quadrature formula and obtain:

$$\overline{g}(s) := \frac{1}{|\gamma(s)|} \int_{\gamma(s)} g(s, \rho_s, \theta)d\theta = g(s, \rho_s, \theta = \pi) + O(|\gamma(s)|^3) \tag{3}$$

Exploiting this quadrature formula in Eq. (2) we get

$$\int_{\Omega} F(t,a)v \simeq \int_{\Lambda_s} |\gamma(s)| f(t,\overline{a}(s))\overline{v}(s)ds. \tag{4}$$

Regarding the definition of the drug release rate, namely $f(t,a)$, we refer to [2, 4, 6]. We assume that the drug release is controlled by drug dissolution and diffusion and we consider a two phase model accounting for solid and dissolved drug. Starting from an initial solid state, drug dissolves and diffuses through the interstices of DES coating substrate in order to finally reach the outer surface and be released. The corresponding model can be analytically approximated by means of asymptotic expansion and the profile of drug release rate $J_s(t,a(t,\mathbf{x}))$ can be explicitly quantified as,

$$J_s(t,a(t,\mathbf{x})) = -(1-a(t,\mathbf{x}))\text{erf}(\Gamma)^{-1}\sqrt{\frac{D_s}{\pi t}} \tag{5}$$

where Γ is a correction coefficient depending on the substrate physical properties and $a(t,\mathbf{x})$ denotes the drug concentration into the artery (referring either to the lumen or the wall). Therefore we can impose:

$$f(t,a) := P(t)(1-a), \qquad \text{with} \qquad P(t) = \text{erf}(\Gamma)^{-1}\sqrt{\frac{D_s}{\pi t}}. \tag{6}$$

and defining

$$\int_{\Omega_a} F_a(t,a)v := 2\pi\rho_s P(t)\int_{\Lambda_s} \xi(s)(1-\overline{a}(s))\overline{v}(s)ds, \tag{7}$$

we obtain a $1D$ model for drug release from a stent. In (7), the function $\xi(s) = |\gamma(s)|/(2\pi\rho_s)$ with $0 \le \xi(s) \le 1$ locally quantifies the fraction of stent surface that is embedded into the artery. By splitting the artery into lumen, Ω_l, and wall, Ω_w, we will later denote by $\xi_l(s)$ and $\xi_w(s)$, respectively, the fractions of the stent surface exposed to each subdomain.

We are now able to define the 3D model that governs the distribution of drug into the artery. We denote by $a_l(t,\mathbf{x})$ and $a_w(t,\mathbf{x})$ the drug concentration for the lumen and the wall, respectively. All drug concentrations are non dimensional values referred to the saturation level of drug dissolved in water. The free drug released to the blood stream obeys to a transport dominated advection-diffusion equation. The drug released into the arterial wall can assume a state where it is dissolved into the plasma permeating the interstices between cells (called *free drug*) or a state where it binds to specific receptors located over proteins that build the extracellular matrix (*bound drug*). The free drug in the arterial wall is able to diffuse and it is transported by the physiological plasma leakage across the arterial wall. Let $b_w(t,\mathbf{x})$ be the density of free receptors with $b_{w,0}(\mathbf{x}) = b_w(t=0,\mathbf{x})$ their initial distribution.

Thanks to mass action law, the rate of change of bound drug is equal to $k_{on}a_w b_w + k_{off}(b_w - b_{w,0})$, where k_{on}, k_{off} are the association and dissociation constants. In conclusion, the release and transport of drug into the artery can be modeled by means of the following equations:

$$\begin{cases} \partial_t a_w - \nabla \cdot (D_w \nabla a_w) + \mathbf{u}_w \cdot \nabla a_w = F_w(t, a_w) + \partial_t b_w & \text{in } \Omega_w \times \mathbb{R}^+ \\ \partial_t b_w + k_{on} a_w b_w + k_{off}(b_w - b_{w,0}) = 0 & \text{in } \Omega_w \times \mathbb{R}^+ \\ \partial_t a_l - \nabla \cdot (D_l \nabla a_l) + \mathbf{u}_l \cdot \nabla a_l = F_l(t, a_l) & \text{in } \Omega_l \times \mathbb{R}^+ \\ \nabla a_w \cdot \mathbf{n}_w = \nabla a_l \cdot \mathbf{n}_l = 0 & \text{on } \Gamma_{cut} \times \mathbb{R}^+ \\ a_l = a_{l,0}, \qquad b_l = 0 & \text{in } \Omega_l \times \{t = 0\} \\ a_w = a_{w,0} \qquad b_w = b_{w,0} & \text{in } \Omega_w \times \{t = 0\} \\ a_l = a_w = 0 & \text{in } \Gamma_{adv} \times \mathbb{R}^+ \end{cases} \tag{8}$$

The vector fields $\mathbf{u}_l$ and $\mathbf{u}_w$ in (8) represent luminal blood flow and transmural plasma filtration velocities respectively. We model blood flow with the incompressible Navier Stokes equations, while plasma filtration is governed by Darcy's model. Since problem (8) features a forcing term of the form $F(t, a) = J_s(t, \overline{a}(t, \mathbf{x}))\delta_{\Lambda_s}$, where δ_s is a Dirac measure, its solution is singular along the line Λ_s. Stability and convergence of a finite element approximation of this problem have been studied in [4] exploiting weighted Sobolev spaces and an augmented formulation.

3 Computational Analysis of Dedicated Stents to Coronary Bifurcations

A model of coronary bifurcation is created with a bifurcation angle of 45°, a thickness of the arterial wall of 0.9 mm and internal diameters of the main branch (MB) and of the side branch (SB) equal to 2.78 and 2.44 mm, respectively. We investigate the adequacy of a stenting procedure where a a dedicated stent is deployed in the side branch (SB), while a standard stent (resembling to a Multilink Vision stent, Abbott Laboratories, Abbott Park, IL; USA) is implanted in the main branch (MB). The dedicated device resembles the Tryton stent (Tryton Medical Inc, Durham, NC, USA), a balloon expandable cobalt-chromium stent specifically designed to be implanted in the SB. The geometry of this bare metal stent (BMS) consists of three distinct regions:

- The main vessel region (proximal) consisting of three long filamentous fronds each joined to circumferential band at the distal and proximal edge of the stent;
- The transition region (central) composed by three panels whose design allows the coverage of the side branch origin;
- The side-branch region (distal) consisting of a standard stent design.

The deformed configuration used to perform the fluid dynamics and drug release analysis was obtained by means of a finite element model aiming to simulate the Inverted Culotte stenting procedure [12]. This technique consists in the expansion of the dedicated Tryton stent in the SB followed by the implantation of the a standard stent in the MB. The procedure is then completed performing a final kissing balloon inflation, the simultaneous expansion of two angioplasty balloons both in the MB and SB. Starting from this deformed configuration, fluid dynamics is achieved by means of a previously developed [5] finite element method coupling the blood flow and intramural plasma filtration in rigid arteries. Blood was modeled as a Newtonian fluid; a steady parabolic velocity profile, whose peak reaches 240 mm/s at the inflow of the vascular district, and 70/30 flow division, at the outflow of the MB and SB respectively, are applied in order to reproduce a physiological coronary mean flow rate over a heartbeat. The drug release and absorption model is complemented with coefficients corresponding to release of heparin. According to [10], the diffusivity of the drug in the arterial wall D_w and in the lumen D_l are set equal to 7.7×10^{-6} mm^2/s and 1.5×10^{-4} mm^2/s. As regards the ligand/receptor interaction involving drug we apply data from [13] where the binding reaction constants are $k_{on} = 10^2\,\mathrm{s}^{-1}$ and $k_{off} = 10^{-2}\,\mathrm{s}^{-1}$ and the average concentration of receptors in the tissue is set to $b_{w,0} = 5$ (we recall that all data refer to non-dimensional concentrations). For the stent struts we assume a circular section equivalent radius equal to $\rho_s = 0.057$ mm. We assume that the initial drug charge, c_s, is ten times the drug saturation level in water. According to the model developed in [4], the value of c_s determines the coefficient Γ that in this case is equal to 0.23.

Two different stenting configurations are investigated: in the first case (labeled as TRY-MB), the stent in the SB is a BMS while in the second case (TRY-MBSB) the device in the SB vessel is a DES. Although the second configuration is not realistic, since a DES version of Tryton stent is not available in commerce, the comparison between the two cases may help to estimate the effectiveness of a dedicated stent in preventing the formation of restenosis into bifurcating vascular districts. Owing to the flexibility of the present model in handling complex stent patterns, we are capable to compare realistic stent deployment with ideal configurations in which all the stent struts are completely apposed to the arterial wall. In such a way, the model can predict the losses of released drug due to strut superposition in case of double stenting techniques, which is a significant issue to be addressed [7].

In order to quantify the drug delivery to the arterial wall, we introduce the dose which is the time-averaged drug concentration at each point of the arterial wall defined by $d_w(x) = T^{-1}\int_0^T \left(c_w(x,t)\,dt\right)$ where $c_w(t,x)$ represents the total drug concentration given by the sum of free $a_w(t,x)$ and bound drug $b_{w,0}-b_w(t,x)$. Drug delivery is analyzed over three specific regions, namely the proximal part of the MB (V_{MBP}), the distal part of the MB (V_{MDP}) and the central part of the SB (V_{SB}) with respect to the bifurcation.

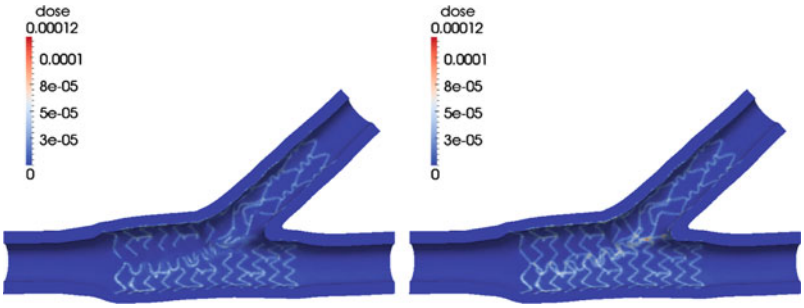


Fig. 2 Contour plots of drug dose in the case of double DES implantation. The case of realistic stent deployment is depicted on the *left*, while the idealized case where both stents are entirely in contact with the artery is reported on the *right*

4 Results and Conclusions

Concerning the comparison between the two stenting configurations TRY-MB and TRY-MBSB, we first observe that either in the real or in idealized cases and in all regions of analysis, drug dosage provided to the artery when two DES are implanted in the MB and in the SB is higher than in the single DES implantation procedure. As expected, overall drug dosage is almost directly proportional to the extension of active surface delivering it. However, the analysis of drug distribution to different arterial regions promotes more interesting considerations.

Concerning drug delivery to the SB, the computational results summarized in Fig. 3 suggest that when the SB is treated with a BMS, drug delivered to the MB hardly reaches the SB. Drug convection from the MB to the SB, which explains the amount of drug accumulated in the SB in this stenting configuration, does not guarantee a significant delivery to the SB. The prescription of a DES to each branch seems to be a more effective option to achieve a more uniform drug distribution to the bifurcation area.

For DES devices originally developed for straight arteries (such as the Multilink Vision one) double stenting procedure is affected by several issues, mainly because two stents are superposed along an extended portion of artery. On the one hand, high amount of metal may lead to toxic drug dosage in the artery. On the other hand, the stent that is superposed to the one implanted first is significantly exposed to washout due to blood flow. For this reason, it is difficult to predict how much of the drug delivered from such stent is actually penetrating into the arterial wall.

Our computational results suggest that the implantation of a dedicated DES in the SB could provide a satisfactory trade off between the aforementioned drawbacks. As illustrated in Fig. 2, the peculiar design of dedicated stents limits the superposition of stent struts to a small portion of arterial wall. For this reason, as shown in Fig. 3, drug delivery to distal parts of the MB is almost equivalent to the case where a DES is deployed in the MB solely. Simultaneously, since only a limited portion of

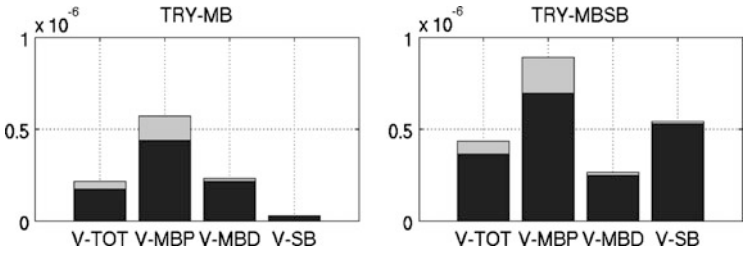


Fig. 3 Dosage (bars) and drug losses with respect to idealized dosage due to stent superposition (incremental bars)

the second DES is detached from the wall, blood washout does not significantly penalize drug delivery.

References

1. A.R. Assali, H.V. Assa, I. Ben-Dor, I. Teplitsky, A. Solodky, D. Brosh, S. Fuchs, and R. Kornowski. Drug-eluting stents in bifurcation lesions: To stent one branch or both? *Catheterization and Cardiovascular Interventions*, 68(6):891–896, 2006.
2. P. Biscari, S. Minisini, D. Pierotti, G. Verzini, and P. Zunino. Controlled release with finite dissolution rate. *SIAM Journal on Applied Mathematics*, 71(3):731–752, 2011.
3. C. Collet, R.A. Costa, and A. Abizaid. Dedicated bifurcation analysis: dedicated devices. *The international journal of cardiovascular imaging*, 27(2):181–188, 2011.
4. C. D'Angelo. *Multiscale modeling of metabolism and transport phenomena in living tissues*. Phd thesis, 2007.
5. C. D'Angelo and P. Zunino. Robust numerical approximation of coupled Stokes' and Darcy's flows applied to vascular hemodynamics and biochemical transport. *ESAIM: Mathematical Modelling and Numerical Analysis*, 45(3):447–476, 2011.
6. C. D'Angelo, P. Zunino, A. Porpora, S. Morlacchi, and F. Migliavacca. Model reduction strategies enable computational analysis of controlled drug release from cardiovascular stents. *SIAM Journal on Applied Mathematics*, 71(6):2312–2333, 2011.
7. V.B. Kolachalama, A.R. Tzafriri, D.Y. Arifin, and E.R. Edelman. Luminal flow patterns dictate arterial drug deposition in stent-based delivery. *Journal of Controlled Release*, 133(1):24–30, 2009.
8. T. Lefevre, Y. Louvard, M.C. Morice, P. Dumas, C. Loubeyre, A. Benslimane, R.K. Premchand, N. Guillard, and J.-F. Piechaud. Stenting of bifurcation lesions: Classification, treatments, and results. *Catheterization and Cardiovascular Interventions*, 49(3):274–283, 2000.
9. W.K. Liu, Y. Liu, D. Farrell, L. Zhang, X.S. Wang, Y. Fukui, N. Patankar, Y. Zhang, C. Bajaj, J. Lee, J. Hong, X. Chen, and H. Hsu. Immersed finite element method and its applications to biological systems. *Computer Methods in Applied Mechanics and Engineering*, 195(13–16):1722–1749, 2006.
10. M.A. Lovich and E.R. Edelman. Computational simulations of local vascular heparin deposition and distribution. *American Journal of Physiology - Heart and Circulatory Physiology*, 40(5):H2014–H2024, 1996.
11. M.C. Morice, P.W. Serruys, J. Eduardo Sousa, J. Fajadet, E.B. Hayashi, M. Perin, A. Colombo, G. Schuler, P. Barragan, G. Guagliumi, and R. Falotico. A randomized comparison of a sirolimus-eluting stent with a standard stent for coronary revascularization. *New England Journal of Medicine*, 346(23):1773–1780, 2002.

12. S. Morlacchi, C. Chiastra, D. Gastaldi, G. Pennati, G. Dubini, and F. Migliavacca. Sequential structural and fluid dynamic numerical simulations of a stented bifurcated coronary artery. *Journal of Biomechanical Engineering*, 133(12):121010, 2011.
13. D.V. Sakharov, L.V. Kalachev, and D.C. Rijken. Numerical simulation of local pharmacokinetics of a drug after intravascular delivery with an eluting stent. *Journal of Drug Targeting*, 10(6):507–513, 2002.
14. P. Zunino, C. D'Angelo, L. Petrini, C. Vergara, C. Capelli, and F. Migliavacca. Numerical simulation of drug eluting coronary stents: Mechanics, fluid dynamics and drug release. *Computer Methods in Applied Mechanics and Engineering*, 198(45–46):3633–3644, 2009.

Coupling Hdiv an H1 Finite Element Approximations for a Poisson Problem

D. de Siqueira, P.R.B. Devloo, and S.M. Gomes

Abstract The purpose of the paper is to approximate an elliptic problem coupling two different formulations. The domain is split into two non-overlapping subdomains. On the first one, the problem is approximated using classical Galerkin method where the primal solution p is searched in H^1 approximation spaces. On the other one, the mixed formulation is applied, which is based on Hdiv and L^2 approximation spaces for the dual ∇p and primal p solutions, respectively. On the interface, the continuity of p and ∇p is imposed strongly, using transmission conditions. The resulting coupled formulation is a saddle point problem, which is solved for high order hierarchical approximation spaces. Numerical simulations for a test problem show consistent rates of convergence when compared with the corresponding classical and mixed formulations in the whole domain.

1 Introduction

In the field of numerical simulation for partial differential equations, there are several techniques named as domain decomposition, with different kinds of target purposes. For instance, they may be designed for parallel implementations or in connection with preconditioning strategies [3].

D. de Siqueira (✉) · S.M. Gomes
IMECC-Unicamp, R. Sérgio Buarque de Holanda 651, 13083-970, Campinas, SP, Brazil
e-mail: denisesiq@gmail.com; soniag@ime.unicamp.br

P.R.B. Devloo
FEC-Unicamp, Av. Albert Einstein 951, 13083-852, Campinas, SP, Brazil
e-mail: phil@fec.unicamp.br

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_44,

The present paper considers domain decomposition for the definition of a coupled formulation for the Poisson problem. The methodology considers the subdivision of the computational domain into two complementary subregions, where different formulations of the same problem are adopted. Namely, a classical Galerkin formulation in one region, and a mixed formulation in its complement, which are coupled by interface transmission conditions.

The coupled formulation is summarized in Sect. 2. For the discretization of this coupled problem, hierarchical high order H^1 and Hdiv approximation spaces are adopted, which have been developed in [2] and [1], respectively. Section 3 contains the simulation results for a test problem, showing consistent rates of convergence when compared with the corresponding classical and mixed formulations on the whole domain.

2 Model Problem and Variational Formulations

On a domain $\Omega \subset I\!R^2$, with Lipschitz boundary $\partial\Omega$, consider the model boundary value problem: To find $(\mathbf{u}, p)$ such that

$$\begin{array}{ll} \mathbf{u} = -\nabla p \text{ in } \Omega, & div(\mathbf{u}) = f \text{ in } \Omega, \\ p = 0 \quad \text{ in } \partial\Omega_D, & \mathbf{u}\cdot\boldsymbol{\eta} = 0 \text{ in } \partial\Omega_N, \end{array} \tag{1}$$

where $f \in L^2(\Omega)$, $\partial\Omega_D$ and $\partial\Omega_N$ are boundary parts where Dirichlet and Neumann conditions are enforced, and $\boldsymbol{\eta}$ denotes the outward unit normal vector to $\partial\Omega$. In a fluid flow context, $\mathbf{u}$ is the flux field, and p is the hydraulic potential (or pressure).

We consider a partition of Ω into two non-overlapping sub-domains Ω_1 and Ω_2 such that $\overline{\Omega} = \overline{\Omega}_1 \cup \overline{\Omega}_2$, $\Gamma = \partial\Omega_1 \cap \partial\Omega_2$, $\partial\Omega_D = \partial\Omega_1 \cap \partial\Omega$, $\partial\Omega_N = \partial\Omega_2 \cap \partial\Omega$. Figure 1 illustrates this kind of domain decomposition,

Let $p_i = p|_{\Omega_i}$ and $\mathbf{u}_i = \mathbf{u}|_{\Omega_i}$, with $i = 1, 2$. We propose to use a classical formulation in Ω_1, and mixed formulation in Ω_2, augmented with boundary transmission conditions. Precisely, consider the set of two coupled problems

$$\begin{cases} -\triangle\, p_1 = f \text{ in } \Omega_1, \\ p_1 = 0 \text{ in } \partial\Omega_D, \\ \nabla p_1 \cdot \boldsymbol{\eta}_1 = -\mathbf{u}_2 \cdot \boldsymbol{\eta}_2 \text{ in } \Gamma. \end{cases} \qquad \begin{cases} \mathbf{u}_2 = -\nabla p_2 \text{ in } \Omega_2, \\ div(\mathbf{u}_2) = f \text{ in } \Omega_2, \\ \mathbf{u}_2 \cdot \boldsymbol{\eta}_2 = 0 \text{ in } \partial\Omega_N, \\ p_1 = p_2 \text{ in } \Gamma. \end{cases}$$

By setting the functional spaces

$$W_1 = \left\{ q \in H^1(\Omega_1) : q|_{\partial\Omega_D} = 0 \right\}, \qquad V_2 = \left\{ \mathbf{v} \in Hdiv(\Omega_2) : \mathbf{v}\cdot\boldsymbol{\eta}|_{\partial\Omega_N} = 0 \right\},$$

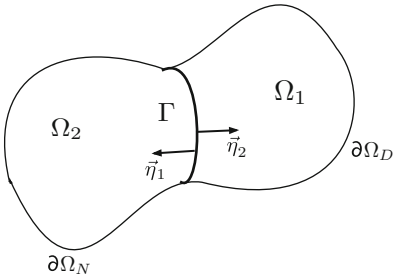


Fig. 1 Domain decomposition

the variational formulation for this coupled problem reads: To find $(p_1, \mathbf{u}_2, p_2) \in W_1 \times V_2 \times L^2(\Omega_2)$ such that

$$\begin{cases} c(p_1, q_1) + c_\Gamma(q_1, \mathbf{u}_2) = f_1(q_1), & \forall\, q_1 \in W_1, \\ a(\mathbf{u}_2, \mathbf{v}_2) + b(\mathbf{v}_2, p_2) - c_\Gamma(p_1, \mathbf{v}_2) = 0, & \forall\, \mathbf{v}_2 \in V_2, \\ b(\mathbf{u}_2, q_2) = f_2(q_2), & \forall\, q_2 \in L^2(\Omega_2), \end{cases} \tag{2}$$

where the bilinear and linear forms are

$$c(p,q) = \int_{\Omega_1} \nabla p \cdot \nabla q \, dx, \qquad a(\mathbf{u}, \mathbf{v}) = \int_{\Omega_2} \mathbf{u} \cdot \mathbf{v} \, dx,$$

$$b(\mathbf{v}, q) = \int_{\Omega_2} q \, div(\mathbf{v}) \, dx, \qquad c_\Gamma(q, \mathbf{v}) = \int_\Gamma q \, (\mathbf{v} \cdot \boldsymbol{\eta}_2) \, ds,$$

$$f_1(q) = \int_{\Omega_1} f \, q \, dx, \qquad f_2(q) = \int_{\Omega_2} f \, q \, dx.$$

This formulation can be interpreted as a saddle point problem, and analysed for stability, existence and uniqueness using Brezzi's theory [4].

Let $\mathscr{T}_{i,h}$ be triangular partions for $\Omega_i, i = 1, 2$, coinciding on the interface Γ. Let $W_h \subset W_1$, $V_h \subset V_2$, and $Y_h \subset L^2(\Omega_2)$ be finite element subspaces. Therefore, the discrete version of problem (2) reads: To find $(p_h^1, \mathbf{u}_h, p_h^2) \in W_h \times V_h \times Y_h$ such that

$$\begin{cases} c(p_h^1, q_h^1) + c_\Gamma(q_h^1, \mathbf{u}_h) = f_1(q_h^1), & \forall\, q_h^1 \in W_h, \\ a(\mathbf{u}_h, \mathbf{v}_h) + b(\mathbf{u}_h, p_h^2) - c_\Gamma(p_h^1, \mathbf{u}_h) = 0, & \forall\, \mathbf{v}_h \in V_h, \\ b(\mathbf{u}_h, q_h^2) = f_2(q_h^2), & \forall\, q_h^2 \in Y_h. \end{cases} \tag{3}$$

The matrix form of problem (3) can be written as

$$\left(\begin{array}{cc:c} A & -(C^\Gamma)^T & B^T \\ C^\Gamma & C & 0 \\ \hdashline B & 0 & 0 \end{array}\right) \left(\begin{array}{c} \mathbf{u}_h \\ p_h^1 \\ p_h^2 \end{array}\right) = \left(\begin{array}{c} 0 \\ f_m^1 \\ f_l^2 \end{array}\right),$$

where the matrices A, C^{Γ}, C and B correspond to the discretization of the bilinear forms $a(\cdot,\cdot)$, $c_{\Gamma}(\cdot,\cdot)$, $c(\cdot,\cdot)$ and $b(\cdot,\cdot)$, respectively, and f^i, $i = 1, 2$, come from the discretization of the linear forms.

3 Numerical Experiments

In order to test the coupled formulation presented in (3), we consider a model problem on the domain $\overline{\Omega} = [0,1] \times [0,1]$, with exact solutions $p(x,y) = \sin \pi\, x \, \sin \pi\, y$ and $\mathbf{u}(x,y) = -[\pi \cos \pi\, x \sin \pi\, y, \pi \cos \pi\, x \sin \pi\, y]^T$, as displayed in Fig. 2.

The domain decomposition $\overline{\Omega} = \overline{\Omega_1} \cup \overline{\Omega_2}$ has $\Omega_1 =]0.5, 1[\times]0, 1[$, and $\Omega_2 =]0, 0.5[\times]0, 1[$, with $\Gamma = \{0.5\} \times [0, 1]$. The meshes $\mathscr{T}_h$, for Ω, and $\mathscr{T}_{i,h}$, for Ω_i, are constructed by triangulations obtained by diagonal subdivision of rectangular uniform meshes in each region, where h denotes the step size in the x-axis.

For the classical formulation in Ω, we take H^1-conform finite element subspaces $W_h \subset H^1(\Omega)$, based on $\mathscr{T}_h$, of type P_k, meaning that the shape functions are polynomials of total degree k. For the mixed formulation in Ω, we consider Hdiv-conform subspaces V_h of type $\mathbf{P}_k$, where the vector shape functions are constructed by the multiplication of H^1-conform shape functions of type P_k by and appropriate vector field to get continuous normal components across element interfaces (see [1]). The subspaces $Y_h \subset L^2(\Omega)$ are formed by piecewise polynomials of total degree $k-1$, such that $div(V_h) = Y_h$. This pair of approximation spaces is named of type $\mathbf{P}_k\ P_{k-1}$.

For the coupled formulation, based on the partitions $\mathscr{T}_{i,h}$, the finite element subspaces $W_h^1 \subset H^1(\Omega_1)$, $V_h^2 \subset Hdiv(\Omega_2)$, and $Y_h^2 \subset L^2(\Omega_2)$ are also of type P_k and $\mathbf{P}_k\ P_{k-1}$, respectively. Figure 3 show the numerical solutions of the coupled formulation for $k = 1$ and $k = 2$, using $h = 1/32$, observing that, in subregion Ω_1, the flux is obtained by post-processing p^1. The solutions in both subregions appear to be consistently coupled accros the interface Γ, and accuracy improvement is noticeable for the higher order approximation.

Figure 4 (left side) displays L^2 errors in pressure for the classical formulation in Ω, to be compared with the errors in p^1, computed in Ω_1, obtained by the coupled formulation. It can be observed that consistent $O(h^{k+1})$ rates of convergence occur in both formulations. On the right side of Fig. 4, the plots are for flux L^2 errors, showing consistent $O(h^k)$ rates of convergence.

Similarly, Fig. 5 (left side) displays L^2 errors in pressure for the mixed formulation in Ω, to be compared with the errors in p^2, computed in Ω_2, as the result of the coupled formulation. Now, the rates of convergence for the pressure decrease to $O(h^k)$, in both formulations. However, the precision in the flux increases to get $O(h^{k+1})$ for the mixed formulation in Ω, as shown on the right side of Fig. 5. However, it can be observed that the convergence of the flux in the coupled formulation, computed in Ω_2, slows down for $k > 1$.

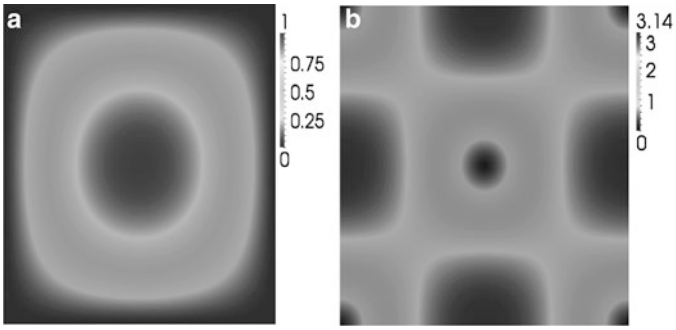


Fig. 2 Pressure (*left*) and flux (*right*) magnitudes of the exact solution

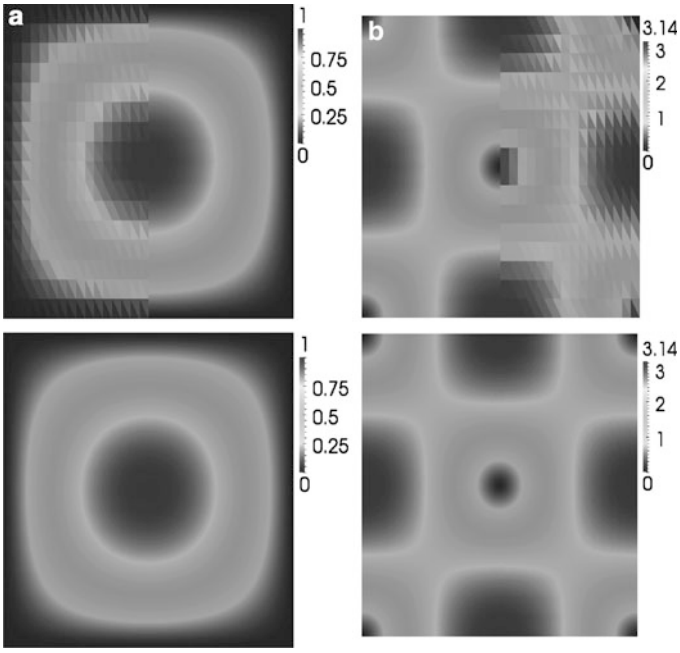


Fig. 3 Pressure (*left*) and flux (*right*) magnitudes of the coupled formulation using approximations of type P_k in Ω_1 and $\mathbf{P}_k$ P_{k-1} in Ω_2, for mesh spacing $h = 1/32$: $k = 1$ (*top*) and $k = 2$ (*bottom*)

This fact can be explained by the influence of the lower order of the flux approximation in Ω_1, affecting the solution in Ω_2 by the action of the matrix C^{Γ}, which components have contributions of both sides. Therefore, to avoid this drawback, we propose to increase the polynomial degree in Ω_1, by taking P_{k+1} type of approximating spaces, in order to balance the coupling with respect to flux

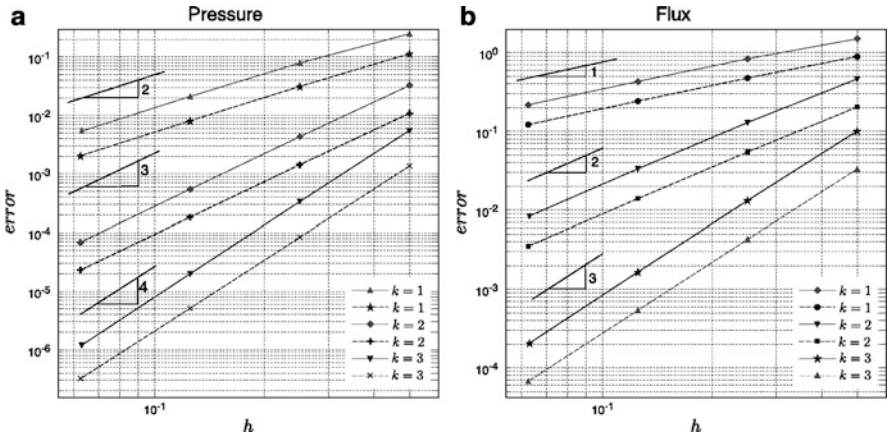


Fig. 4 Errors in pressure (**a**) and flux (**b**) approximations, using classical formulation of type $\mathbf{P}_k$ (*solid*, measured in Ω) and coupled formulation of type $\mathbf{P}_k$ in Ω_1 and $\mathbf{P}_k$ P_{k-1} in Ω_2 (*dashed*, measured in Ω_1)

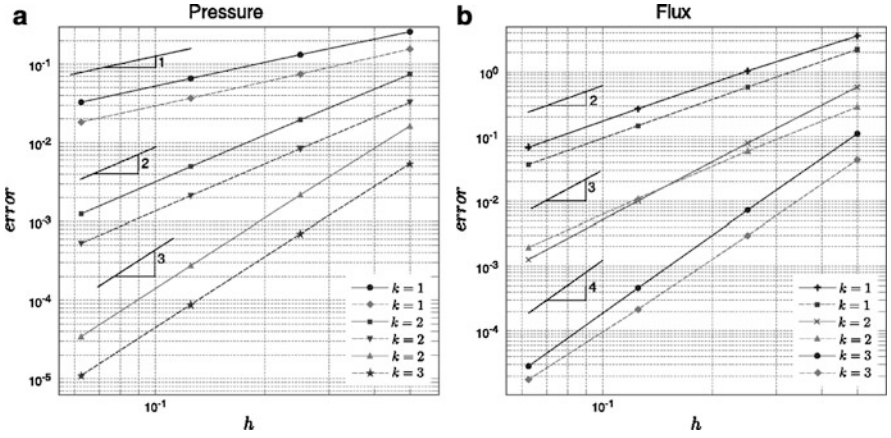


Fig. 5 Errors in pressure (**a**) and flux (**b**) approximations, using mixed formulation of type $\mathbf{P}_k$ P_{k-1} (*solid*, measured in Ω) and coupled formulation of type P_k in Ω_1 and $\mathbf{P}_k$ P_{k-1} in Ω_2 (*dashed*, measured in Ω_2)

approximations. Figure 6 show the results after this enrichment for $k = 2$, where the same $O(h^3)$ rate of convergence is obtained in the flux computed by the mixed formulation in Ω and by the coupled formulation, measured in Ω_2.

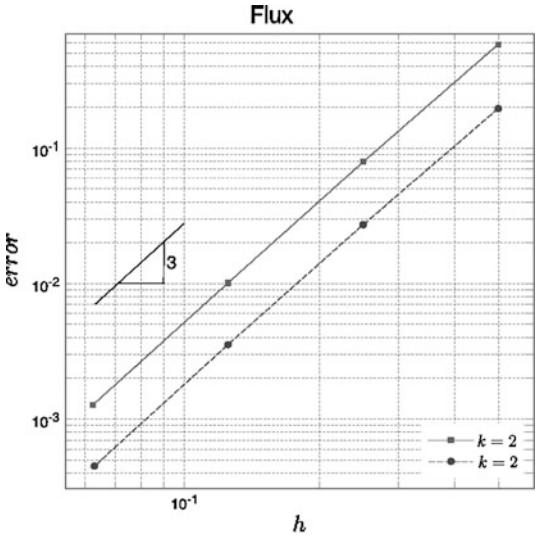


Fig. 6 Errors in flux approximations, using mixed formulation of type $\mathbf{P}_2\ P_1$ (*solid*, measured in Ω) and coupled formulation of type P_3 in Ω_1 and $\mathbf{P}_2\ P_1$ in Ω_2 (*dashed*, measured in Ω_2)

Acknowledgements The authors thankfully acknowledges financial support from the Brazilian National Agency of Petroleum, Natural Gas and Biofuels (ANP – PETROBRAS). P. Devloo and S.M. Gomes thankfully acknowledges financial support from CNPq – the Brazilian Research Council.

References

1. Siqueira, D., Devloo, P. R. B., and Gomes, S.M.: Hierarchical high order finite element approximation spaces for Hdiv and Hcurl. In: Kreiss, G., Lötstedt, P., Malqvist, A. Neytcheva, M. (eds.) Numerical Mathematics and Advanced Applications, pp. 267–274. Springer, Uppsala (2009)
2. Bravo, C. M. A. A., Devloo, P. R. B., and Rylo, E.C.: Systematic and generic construction of shape functions for p-adaptive meshes of multidimensional finite elements. Comput. Meth. Appl. Mech. Eng. **198**, 1716–1725 (2009)
3. Quarteroni, A., and Valli, A.: Domain Decomposition Methods for Partial Differential Equations, Oxford Science Publications (1999)
4. Brezzi, F., and Fortin, M.: Mixed and Hybrid Finite Element Methods, Springer (1991)

Nodal Interpolation Between First-Order Finite Element Spaces in 1D is Uniformly H^1-Stable

T. Dickopf

Abstract This paper is about the stability w.r.t. the H^1-semi-norm of the nodal interpolation operator acting between non-nested finite element spaces. (An earlier, slightly less general version of the main result has been proved in the author's thesis (Dickopf, Multilevel methods based on non-nested meshes. Ph.D. thesis, University of Bonn, 2010. http://hss.ulb.uni-bonn.de/2010/2365, Chap. 5.1). Lively and fruitful discussions first during the ENUMATH conference in September and then during the Söllerhaus Workshop on Domain Decomposition Methods in October 2011 have encouraged the author to rework the analysis of the nodal interpolation over intervals and present it in this extended and considerably revised form.) We show that, for arbitrary spaces of piecewise linear functions of one variable, the H^1-stability constant is bounded by one without any assumptions on the mesh sizes or on the relations between the meshes. We also give counterexamples for the nodal interpolation in higher order finite element spaces.

1 Introduction

In the numerical analysis of the finite element method, the existence of families of stable approximation operators is exploited frequently. Applications range from a priori and a posteriori error estimates over the analysis of coupled discretizations to convergence proofs of iterative solvers; see [1–9] and their references. The perhaps most influential work on quasi-interpolation in finite element spaces is [4]; see [1,5] for more recent overviews. However, nodal interpolation certainly plays a prominent role as it is by far the simplest operator both conceptionally and computationally.

T. Dickopf (✉)
Institute of Computational Science, University of Lugano
Via G. Buffi 13 6904, Lugano, Switzerland
e-mail: thomas.dickopf@usi.ch

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_45,

We consider the nodal interpolation operator between finite element spaces associated with non-nested meshes. Applications in this particular setting can be found in [2,3,5,6,9]. While several proofs of the H^1-stability have been put forward in the 1990s [2,3,9], driven by the research of domain decomposition methods with non-nested coarse spaces, less is known about the size of the constants.

We show a uniform H^1-stability estimate for the nodal interpolation between spaces of piecewise linear functions of one variable. Neither the mesh sizes nor rela- tions between the meshes enter the constants. This further improves the understanding of the interpolation operator in this setting. The size of the stability constants is of particular interest for the use of iterated operators, which appear, e.g., in [5,7].

2 Operators Between Unrelated Finite Element Spaces

In this section, we describe the problem setting for the analysis of operators between finite element spaces associated with non-nested meshes. First, let us recall several standard notations from functional analysis.

For a bounded connected open set (an interval) $\Omega \subset \mathbb{R}$, let $L^2(\Omega)$ be the Hilbert space of square integrable functions in Ω with inner product $(v,w)_{L^2(\Omega)} := \int_\Omega vw\,dx$ and norm $\|\cdot\|_{L^2(\Omega)} := (\cdot,\cdot)^{1/2}_{L^2(\Omega)}$. The symbol $L^\infty(\Omega)$ represents the space of essentially bounded functions with norm $\|v\|_{L^\infty(\Omega)} := \operatorname{ess\,sup}_{x\in\Omega}|v(x)|$. By $H^1(\Omega)$, as customary, we denote the Sobolev space of functions with square integrable weak derivatives in the domain Ω. The corresponding norm and seminorm are $\|\cdot\|_{H^1(\Omega)}$ and $|\cdot|_{H^1(\Omega)}$, respectively. Moreover, the subspace with vanishing image of the usual trace operator to the boundary $\partial\Omega$ is called $H^1_0(\Omega)$.

This paper analyzes interpolation operators from one finite element space to another. The considered finite element spaces are associated with non-nested meshes possibly representing different domains; they are denoted by X and Y, respectively. In the following, we introduce the basic notations for the symbol $A \in \{X, Y\}$.

Let $\mathscr{T}_A$ be a mesh of the domain $\Omega_A \subset \mathbb{R}$, i.e., a non-overlapping decomposition into finitely many intervals $T \in \mathscr{T}_A$ such that $\Omega_A = \operatorname{int}\left(\cup_{T\in\mathscr{T}_A}\overline{T}\right)$. As usual, local mesh size functions $h_A \in L^\infty_>(\Omega_A) := \{v \in L^\infty(\Omega_A) \mid \exists\, \alpha > 0,\ \text{such that } v(x) > \alpha \text{ for a.e. } x \in \Omega_A\}$ are introduced for instance defined a.e. by $h_A(x) := \operatorname{diam}(T)$ if $x \in T$. When thinking of families of meshes with decreasing discretization parameter, one notes that no shape regularity assumption needs to be considered in the 1D case. Moreover, we do not assume quasi-uniformity in this paper.

We denote the set of nodes of $\mathscr{T}_A$ by $\mathscr{N}_A$ and abbreviate $n_A := |\mathscr{N}_A|$. Then, the symbol A denotes the space of first-order conforming finite elements; its nodal (Lagrange) basis is $\Lambda_A = (\lambda^A_p)_{p\in\mathscr{N}_A}$ with $\lambda^A_p(q) = \delta_{pq}$, $p, q \in \mathscr{N}_A$.

In many applications, approximation operators are employed with the following two asymptotic properties for decreasing mesh size of the target finite element space. In fact, examples can be found in every single of the references [1–9].

Definition 1. Let $\Omega_Y \subset \Omega_X$ be domains as above. Given a subspace $X \subset H^1(\Omega_X)$ and a finite element space $Y \subset H^1(\Omega_Y)$ with discretization parameter $h_Y \in L^\infty_>(\Omega_Y)$, an operator $\Pi : X \to Y$ is called H^1-stable in X if there is a constant C_{stab} independent of h_Y such that

$$|\Pi v|_{H^1(\Omega_Y)} \leq C_{\text{stab}} |v|_{H^1(\Omega_X)}, \quad \forall\, v \in X.$$

The operator $\Pi : X \to Y$ is said to satisfy the L^2-approximation property if

$$\|h_Y^{-1}(v - \Pi v)\|_{L^2(\Omega_Y)} \leq C_{\text{approx}} |v|_{H^1(\Omega_X)}, \quad \forall\, v \in X,$$

with C_{approx} independent of h_Y.

Here, no relation between X and Y has been specified other than the fact that functions from X are also well-defined in the domain Ω_Y. In the present context, X is a finite element space as described above. Note that this is not prerequisite for the definition to be applicable, though. The described properties are well-known; they are relevant for both analysis and practical computations.

The Nodal Interpolation Operator

Let us now consider the standard finite element interpolation. For an exhaustive review of various approximation operators, locally and globally defined mappings between finite element spaces, we refer to [5]. The nodal interpolation operator, which maps continuous functions to the finite element space Y, is defined by

$$\mathscr{I} : \mathscr{C}^0(\Omega) \to Y, \quad v \mapsto \mathscr{I} v := \sum\nolimits_{p \in \mathscr{N}_Y} v(p) \lambda_p^Y. \tag{1}$$

Assume Ω is such that $\mathscr{I}$ is well-defined. Evidently, the operator is surjective, namely $\mathscr{I}(\mathscr{C}^0(\Omega)) = Y$, and a projection, i.e., for any $v \in \mathscr{C}^0(\Omega)$ it is $\mathscr{I}\mathscr{I} v = \mathscr{I} v$.

From a computational point of view, nodal interpolation is very attractive. With one function evaluation per basis function in Λ_Y, it is without any doubt the cheapest way to transfer information to a finite element space in a reasonable manner.

For shape regular meshes (in 1D, 2D and 3D), the operator $\mathscr{I}$ possesses the H^1-stability and L^2-approximation properties of Definition 1 when restricted to finite element spaces. In the literature, several different proofs have been brought forth; see [2,3,9]. One notes that the stability estimates usually depend on the shape regularity of the meshes which, in general, leads to $C_{\text{stab}} \gg 1$. In contrast, we prove that the H^1-stability constant C_{stab} is uniformly bounded by one if $\mathscr{I}$ maps between unrelated first-order finite element spaces in 1D.

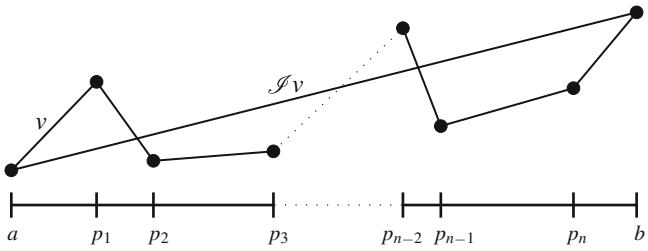


Fig. 1 Illustration of the notations for the interpolation operator $\mathscr{I} : S_n(I, \boldsymbol{p}) \to S_0(I)$

3 Uniform Boundedness of the Linear Nodal Interpolation in 1D

In this section, we show that the nodal interpolation operator between first-order finite element spaces over intervals is uniformly bounded w.r.t. the H^1-semi-norm. More precisely, $C_{\text{stab}} \leq 1$ independent of the choice of the non-nested meshes.

Before elaborating two lemmas, we introduce some notations regarding spaces of piecewise linear functions. For an interval $I = [a, b]$ and $n > 0$, let $\boldsymbol{p} = (p_i)_{i=1,\dots,n}$ be a vector of intermediate points with $a < p_1 < \ldots < p_n < b$ forming the subintervals $I_1 = [a, p_1]$, $I_i = [p_{i-1}, p_i]$ for $2 \leq i \leq n$, and $I_{n+1} = [p_n, b]$. Then, the space of continuous piecewise first-order polynomials is denoted by

$$S_n(I, \boldsymbol{p}) := \left\{ v \in \mathscr{C}^0(I) \mid v|_{I_i} \in \mathbb{P}_1(I_i) \text{ for } 1 \leq i \leq n+1 \right\}. \tag{2}$$

To have a consistent notation for first-order finite element spaces with $n \geq 0$ interior mesh points, we write $S_0(I) := \mathbb{P}_1(I)$.

In the following auxiliary results (Lemmas 1 and 2), we consider the linear interpolation on the whole interval, which only takes the values at the endpoints a and b into account. This operator is denoted by $\mathscr{I} : S_n(I, \boldsymbol{p}) \to S_0(I)$; see Fig. 1 for an illustration.

We first prove that the linear interpolant has minimal H^1-semi-norm among all piecewise linear functions with only one intermediate point; in other words, the operator $\mathscr{I} : S_1(I, \boldsymbol{p}) \to S_0(I)$ is H^1-stable with $C_{\text{stab}} \leq 1$.

Lemma 1. *Let $I = [a, b]$ be an interval and $\boldsymbol{p} = (p_1)$ be an intermediate point, i.e., $a < p_1 < b$. Then, the interpolation operator $\mathscr{I} : S_1(I, \boldsymbol{p}) \to S_0(I)$ satisfies*

$$|\mathscr{I}v|_{H^1(I)} \leq |v|_{H^1(I)}, \quad \forall\, v \in S_1(I, \boldsymbol{p}). \tag{3}$$

Proof. Consider a function $v \in S_1(I, \boldsymbol{p})$. Without loss of generality, let $I = (0, 1)$ and then $v(0) = 0$ and $v(1) = 1$. This leaves as variables the coordinate $p := p_1 \in (0, 1)$ and the intermediate value $u := v(p) \in \mathbb{R}$; see Fig. 2. We introduce the functional

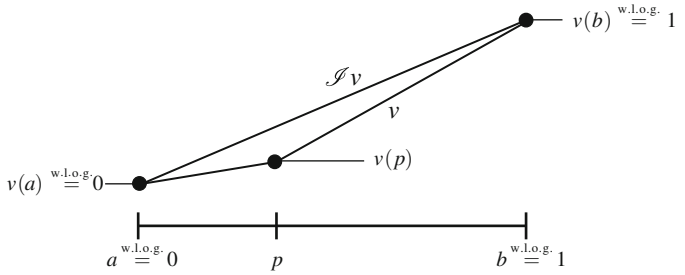


Fig. 2 Notations in the proof of Lemma 1 for the interpolation operator $\mathscr{I} : S_1(I, \boldsymbol{p}) \to S_0(I)$

$$\mathscr{H} : (0,1) \times \mathbb{R} \to \mathbb{R}, \quad \mathscr{H}(p,u) := \frac{u^2}{p} + \frac{(1-u)^2}{1-p}$$

describing the square of the H^1-semi-norm of the piecewise linear function connecting the three points $(0,0)$, (p,u) and $(1,1)$. It is easy to see that, given any $p \in (0,1)$, the functional $\mathscr{H}$ attains its minimum if the measured function is linear, namely if $u = p$. Indeed, the calculation of the first partial derivatives shows that the gradient $\nabla\mathscr{H}(p,u)$ vanishes if and only if $u = p$. In addition, the Hessian at these points

$$\nabla^2\mathscr{H}(p,p) = \frac{2}{p(1-p)} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

is positive semi-definite. Therefore, the functional $\mathscr{H}$ has a set of minima (with value 1) on the diagonal line $\{p = u\}$. Evidently, $\mathscr{H}$ cannot be minimal at the boundary of its domain. This implies (3). An alternative proof can be obtained.[1] □

The second lemma extends the result to the linear interpolation of a piecewise linear function with $n > 1$ intermediate points. This is the situation that has been illustrated earlier in Fig. 1.

Lemma 2. *Let $n \geq 1$ and $\boldsymbol{p} = (p_i)_{i=1,\ldots,n}$ be intermediate points in an interval $I = [a,b]$, i.e., $a < p_1 < \ldots < p_n < b$. Then, the operator $\mathscr{I} : S_n(I, \boldsymbol{p}) \to S_0(I)$ satisfies*

$$|\mathscr{I}v|_{H^1(I)} \leq |v|_{H^1(I)}, \quad \forall\, v \in S_n(I, \boldsymbol{p}). \tag{4}$$

Proof. The case $n = 1$ is Lemma 1. Let $n > 1$ and $v \in S_n(I, \boldsymbol{p})$ arbitrary. Assume that the assertion (4) holds true for $n-1$. By restriction, we may consider v as element of $S_{n-1}(I', \boldsymbol{p}')$ associated with the smaller interval $I' := [p_1, b]$ and the intermediate points $\boldsymbol{p}' := (p_2, \ldots, p_n)$. By assumption, we have $|\mathscr{I}'v|_{H^1(I')} \leq |v|_{H^1(I')}$ for the operator[2] $\mathscr{I}' : S_{n-1}(I', \boldsymbol{p}') \to S_0(I')$.

[1] Exploit the fact that for $p \in (0,1)$ one has $u^2/p + (1-u)^2/(1-p) = 1 + (u-p)^2/(p(1-p)) \geq 1$.

[2] For the symbol $\mathscr{I}$ introduced in (1), it is usually not necessary to mark the dependence of the interpolation operator on the considered spaces. We do this here to clarify the recursive argument.

Let $\boldsymbol{p}'' := (p_1)$ and $w \in S_1(I, \boldsymbol{p}'')$ be defined piecewise by v in $I \setminus I'$ and $\mathscr{I}'v$ in I'. In particular, $\mathscr{I}w = \mathscr{I}v$. Lemma 1 applied to this situation implies

$$|\mathscr{I}v|^2_{H^1(I)} = |\mathscr{I}w|^2_{H^1(I)} \leq |w|^2_{H^1(I)} = |v|^2_{H^1(I\setminus I')} + |\mathscr{I}'v|^2_{H^1(I')}$$

$$\leq |v|^2_{H^1(I\setminus I')} + |v|^2_{H^1(I')} = |v|^2_{H^1(I)}.$$

This concludes the proof by induction. □

Of course a direct proof similar to the one of Lemma 1 is possible, too.

Let us proceed with the main result about the stability of the nodal interpolation. For this purpose, we go back to the general case of two non-nested finite element spaces, which has been described in Sect. 2. The following theorem yields the desired uniform stability estimate without any assumptions on the mesh sizes or on the relations between the meshes.

Theorem 1. *The nodal interpolation operator $\mathscr{I} : X \to Y$ satisfies the following uniform H^1-stability estimates:*
If $\Omega_Y \subset \Omega_X$,

$$|\mathscr{I}v|_{H^1(\Omega_Y)} \leq |v|_{H^1(\Omega_Y)} \leq |v|_{H^1(\Omega_X)}, \quad \forall\, v \in X, \tag{5}$$

otherwise,

$$|\mathscr{I}\mathscr{E}v|_{H^1(\Omega_Y)} \leq |v|_{H^1(\Omega_Y\cap\Omega_X)}, \quad \forall\, v \in X,\ v|_{\partial\Omega_X} = 0, \tag{6}$$

where $\mathscr{E} : H^1_0(\Omega_X) \to H^1(\Omega_X \cup \Omega_Y)$ is the natural extension by zero.

We emphasize that the symbols $\subset$ and $\supset$ always include the case of equality. In addition, note that the restriction to functions in $H^1_0(\Omega_X)$ in (6) ensures that their extensions by zero are weakly differentiable.

Proof. First, consider the case $\Omega_Y \subset \Omega_X$. Let $v \in X$ be arbitrary. If $T_Y \in \mathscr{T}_Y$ is completely contained in an element of $\mathscr{T}_X$, we have locally exact interpolation, i.e., $(\mathscr{I}v)|_{T_Y} = v|_{T_Y}$; thus, $|\mathscr{I}v|^2_{H^1(T_Y)} = |v|^2_{H^1(T_Y)}$.

Now, let $T_Y \in \mathscr{T}_Y$ be an element which has intersections with $n+1 > 1$ elements in $\mathscr{T}_X$. (Recall that elements are open intervals.) This means that there is a vector $\boldsymbol{p}$ of n intermediate points such that $v|_{T_Y} \in S_n(T_Y, \boldsymbol{p})$. Therefore, Lemma 2 is applicable for $\mathscr{I} : S_n(T_Y, \boldsymbol{p}) \to S_0(T_Y)$ and yields $|\mathscr{I}v|^2_{H^1(T_Y)} \leq |v|^2_{H^1(T_Y)}$. See Fig. 3 for an illustration of the applicability. Summation over all $T_Y \in \mathscr{T}_Y$ concludes the proof of (5).

For the second assertion (6), we realize that the nodal interpolation is identically zero on the exterior elements $T_Y \subset \Omega_Y \setminus \Omega_X$. On the interior elements $T_Y \subset \Omega_X$, the local estimates hold as before. We now consider the remaining cases. Let $\Omega_Y \not\subset \Omega_X$ and assume that $T_Y \cap \partial\Omega_X \neq \emptyset$. In this situation, Lemma 2 yields the desired local estimate, too, either with $v(a) = v(p_1) = 0$ (left boundary) or with $v(p_n) = v(b) = 0$ (right boundary). Consequently, the estimate (6) follows as before by adding up the local contributions. □

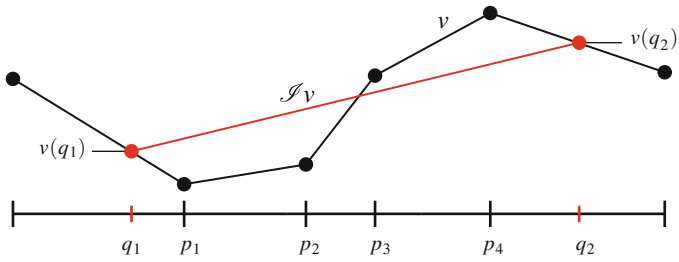


Fig. 3 Illustration of the applicability of Lemma 2 on the element $T_Y \in \mathscr{T}_Y$ with endpoints q_1 and q_2 in the proof of Theorem 1. Here, we exemplarily show the case $n = 4$, i.e., T_Y intersects five elements in $\mathscr{T}_X$ and the intermediate points are $p_1, \ldots, p_4$

Let us finally consider the interpolation operator $\mathscr{I}_0 : X \to Y \cap H_0^1(\Omega_Y)$ enforcing zero function values at $\partial\Omega_Y$, i.e., $(\mathscr{I}_0 v)(p) = 0$ if $p \in \mathscr{N}_Y \cap \partial\Omega_Y$. We emphasize that the previous results do generally not hold for this mapping unless restricted to functions already satisfying the boundary conditions.

Theorem 2. *The interpolation operator $\mathscr{I}_0 : X \to Y \cap H_0^1(\Omega_Y)$ satisfies*

$$|\mathscr{I}_0 \mathscr{E} v|_{H^1(\Omega_Y)} \le |v|_{H^1(\Omega_Y \cap \Omega_X)}, \quad \forall\, v \in X,\ v|_{\partial\Omega_X} = 0,\ v|_{\partial\Omega_Y \cap \overline{\Omega}_X} = 0, \tag{7}$$

where $\mathscr{E} : H_0^1(\Omega_X) \to H^1(\Omega_X \cup \Omega_Y)$ is the natural extension by zero.

Proof. The input v specified by (7) vanishes on the elements $T_X \in \mathscr{T}_X$ with $T_X \cap \partial\Omega_Y \neq \emptyset$ or $T_X \subset \Omega_X \setminus \Omega_Y$, if any. This implies that $\mathscr{I}_0$ coincides with the standard operator on the considered subspace because $(\mathscr{I} v)(p) = 0$ if $p \in \mathscr{N}_Y \cap \partial\Omega_Y$. Therefore, the assertion follows immediately from the previous theorem. □

It seems not feasible to use the same elementary but elegant techniques to prove analogous results for unstructured meshes in 2D or 3D. This is because the constant of a shape regularity assumption comes into play, on the one hand, by a local inverse inequality and, on the other hand, by the sum over neighboring elements.

4 Counterexamples for Higher Order Nodal Interpolation

In this section, we present counterexamples showing that the H^1-stability constants can generally not be bounded by one for the nodal interpolation in higher order finite element spaces.

In analogy to (2), let the spaces of continuous piecewise second- and third-order polynomials be denoted by $S_n^Q(I, \boldsymbol{p})$ and $S_n^C(I, \boldsymbol{p})$, respectively. Consider an interval $T_Y = (0, 1)$ with two decompositions, first into $T_v^1 = (0, 1/2)$ and $T_v^2 = (1/2, 1)$ and then into $T_w^1 = (0, 1/3)$ and $T_w^2 = (1/3, 1)$, respectively. Then, the piecewise linear functions v and w defined by

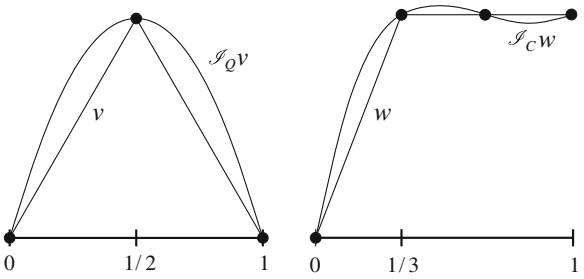


Fig. 4 Quadratic (*left*) and cubic (*right*) nodal interpolation in the element $T_Y = (0, 1)$ of the input functions v and w that are piecewise linear w.r.t. the decompositions with $1/2$ and $1/3$ as intermediate points, respectively

$$v(x) = \begin{cases} 2x, & x \in \overline{T}_v^1 \\ 2 - 2x, & x \in \overline{T}_v^2 \end{cases}, \qquad w(x) = \begin{cases} 3x, & x \in \overline{T}_w^1 \\ 1, & x \in \overline{T}_w^2 \end{cases}$$

obviously satisfy $v \in S_1^Q(T_Y, 1/2)$ and $w \in S_1^C(T_Y, 1/3)$. For these functions, it is easy to evaluate the quadratic interpolation $\mathscr{I}_Q : S_1^Q(T_Y, 1/2) \to S_0^Q(T_Y)$ and the cubic interpolation $\mathscr{I}_C : S_1^C(T_Y, 1/3) \to S_0^C(T_Y)$. We have $\mathscr{I}_Q v(x) = -4x^2 + 4x$ and $\mathscr{I}_C w(x) = 4.5x^3 - 9x^2 + 5.5x$ for all $x \in T_Y$. This is illustrated in Fig. 4. Finally, straightforward calculations of the H^1-semi-norms show that

$$|\mathscr{I}_Q v|_{H^1(T_Y)} / |v|_{H^1(T_Y)} = 2/\sqrt{3} > 1$$

and

$$|\mathscr{I}_C w|_{H^1(T_Y)} / |w|_{H^1(T_Y)} = \sqrt{37/30} > 1.$$

Evidently, one does not need to choose piecewise linear input functions; this has been done only for the sake of simplicity. Moreover, note that counterexamples for fourth-order and higher can be constructed in a similar fashion.

Acknowledgements This work was partly supported by the Iniziativa Ticino in Rete. The author is very grateful to Rolf Krause for his ongoing support during the last years. Thanks to an anonymous referee for hinting at the alternative proof of Lemma 1.

References

1. T. Apel. Interpolation in h-version finite element spaces. In E. Stein et al., editor, *Encyclopedia of Computational Mechanics. Vol. 1. Fundamentals*, pages 55–72. Wiley, 2004.
2. X. Cai. The use of pointwise interpolation in domain decomposition methods with non-nested meshes. *SIAM J. Sci. Comput.*, 16(1):250–256, 1995.

3. T. Chan, B. Smith, and J. Zou. Overlapping Schwarz methods on unstructured meshes using non-matching coarse grids. *Numer. Math.*, 73(2):149–167, 1996.
4. P. Clément. Approximation by finite element functions using local regularization. *RAIRO Anal. Numér.*, 9(R-2):77–84, 1975.
5. T. Dickopf. *Multilevel Methods Based on Non-Nested Meshes*. PhD thesis, University of Bonn, 2010. http://hss.ulb.uni-bonn.de/2010/2365.
6. H. H. Kim and O. B. Widlund. Two-level Schwarz algorithms with overlapping subregions for mortar finite elements. *SIAM J. Numer. Anal.*, 44(4):1514–1534, 2006.
7. P. Oswald. Intergrid transfer operators and multilevel preconditioners for nonconforming discretizations. *Appl. Numer. Math.*, 23(1):139–158, 1997.
8. O. Steinbach. *Stability Estimates for Hybrid Coupled Domain Decomposition Methods*, volume 1809 of *Lecture Notes in Mathematics*. Springer, Berlin, 2003.
9. A. Toselli and O. Widlund. *Domain Decomposition Methods – Algorithms and Theory*, volume 34 of *Springer Ser. Comput. Math.* Springer, 2005.

M-Adaptation Method for Acoustic Wave Equation on Rectangular Meshes

V. Gyrya and K. Lipnikov

Abstract A novel discretization strategy, dubbed m-adaptation, is developed for solving the acoustic wave equation (in the time domain) on rectangular meshes. The developed method is based on a scheme that is second-order accurate in space and time but has the sixth-order numerical anisotropy on square meshes and the fourth order dispersion on rectangular meshes.

1 Introduction

Let us consider a 2D acoustic wave equation in a homogeneous medium. Its semi-discrete form is

$$M u_{tt} = Au, \tag{1}$$

where M and A are the mass and stiffness matrices, respectively. Our focus is on explicit schemes and long integration time. For this, we have to resolve two numerical issues: efficient inversion of the matrix M on each time step and small numerical dispersion.

Finite difference (FD) schemes use a diagonal mass matrix but have a relatively high numerical dispersion for low-order schemes. The later is typically countered by increasing either the mesh density or the stencil size and the discretization order (see e.g. [1]). Another approach is to combine high-order time integration schemes (see e.g. [2]) with discretization methods optimized in space (see e.g. [3–5]). For instance, in [6, 7], an optimized 9-point FD method in the frequency domain is derived using a weighted average of the standard and a "diagonal" FD schemes. In [8], this idea is extended to three-dimensions.

V. Gyrya (✉) · K. Lipnikov
Los Alamos National Laboratory, MS B284, Los Alamos, NM, 87545, USA
e-mail: vitaliy_gyrya@lanl.gov; lipnikov@lanl.gov

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_46,

In a finite element (FE) method, the mass matrix is often lumped which does not change the formal order of accuracy but leads to a severe numerical dispersion [9]. A similar effect is observed for the reproducing kernel particle method [10] and the natural element method [11]. The spectral element method based on the Gauss-Lobatto quadrature points results in a diagonal matrix M, but is mainly restricted to quadrilateral meshes [12, 13]. Thus, resolving complex geometries using this method represents a challenge. The discontinuous Galerkin FE and extended FE methods work on general meshes but lead to large algebraic problems [14, 15].

Our approach is based on replacing the mass matrix by the matrix $DM^{-1}D$, where D is the lumped mass matrix and M is the original matrix. This approach was considered in [16, 17]. In [17], it was combined with special quadratures for calculating stiffness and mass matrices on square elements. The resulting scheme being formally the second-order accurate in space and time exhibits the fourth-order numerical dispersion.

The numerical quadratures in [17] have minimal accuracy sufficient for preserving the order of discretization error. This implies that they may be not unique and new schemes with better properties may be derived. Mathematically, the family of such schemes can be formalized using the technology developed for mimetic finite difference (MFD) schemes. In these schemes, the mass and stiffness matrices are derived from elemental consistency conditions. The number of these conditions is related to the accuracy of the numerical scheme is often less than the size of elemental matrices. The MFD technology provides a parameterized family of acceptable elemental mass and stiffness matrices. Analysis of this family on square meshes allows one to obtain a scheme which is second-order accurate in space and time but has the fourth-order numerical dispersion and the six-order numerical anisotropy [18]. Here, the analysis is extended to rectangular meshes. We show existence of a two-parameter subfamily of schemes with the fourth-order numerical dispersion and anisotropy.

The theory of MFD methods states that the dimension of a parametric family of admissible elemental mass and stiffness matrices grows quadratically with the discretization order and the number of vertices in a mesh element. For example, for a second-order scheme on a hexahedral mesh, the number of parameters in each computational cell reaches 38. Many existing schemes belong to this family [18].

The paper outline is as follows. In Sect. 2, we introduce briefly the MFD family of schemes. In Sect. 3, we perform dispersion analysis and find subfamily of schemes that minimize numerical dispersion and numerical anisotropy. In Sect. 4, we illustrate our funding with numerical experiments and comment on stability range of the derived schemes.

2 Second-Order MFD Scheme on Quadrilaterals

Consider the scalar wave equation

$$u_{tt} = c^2 \triangle u + f \qquad \text{in} \quad \Omega, \tag{2}$$

where c is the wave velocity, $u(\mathbf{x}; t)$ and $f(\mathbf{x}; t)$ are scalar displacement and force, respectively. The system is closed by selecting appropriate initial and boundary conditions. For simplicity, we consider homogeneous Dirichlet boundary conditions. The weak form of Eq. (2) is

$$\int_\Omega u_{tt} v \, dV = -\int_\Omega c^2 \nabla u \cdot \nabla v \, dV + \int_\Omega f v \, dV \qquad \forall v \in H_0^1(\Omega). \tag{3}$$

In the MFD scheme, both integrals are calculated element-by-element. First, for each elemental matrix, a set of consistency conditions is derived. These conditions form a system of algebraic equations with respect to an unknown matrix. Second, a parametric solution of this system is obtained.

Let Ω_h denote a conformal mesh with quadrilateral elements E. We denote edges of E by e_k^E and its vertices by a_k^E, $k = 1, 2, 3, 4$. Let $|E|$ and $|e_k^E|$ denote area of element E and length of edge e_k^E, respectively. Finally, let $\mathbf{n}_k^E$ be the exterior unit normal vector to edge e_k^E. Since most of our considerations are performed on a single element, we omit hereafter the superscript E.

We select discrete unknowns at mesh vertices. Let U_E be the vector of discrete unknowns, u_{a_k}, associated with the vertices a_k of element E. For a quadrilateral, this is a vector of size four.

2.1 Elemental Stiffness Matrix

Consider a quadrilateral E. The elemental stiffness matrix A_E is a symmetric 4×4 semi-positive definite matrix that approximates the energy integral:

$$V_E^T \, A_E \, U_E \approx \int_E c^2 \nabla v_h \cdot \nabla u_h \, dV. \tag{4}$$

In the MFD methods, the shapes of functions v_h and u_h do not come into a play. Instead, these functions are defined in a non-unique way via a few properties. First, their values at mesh vertices are exactly our degrees of freedom. Second, they can be integrated exactly on each edge using the trapezoidal quadrature rule. Third, the convergence theory imposes mesh-independent bounds on their energy norms [19].

The first condition for the matrix A_E is the P_1-compatibility, i.e. Eq. (4) becomes the identity when u_h is a linear function. There are only three linearly independent linear functions: $\phi_1 = 1$, $\phi_2 = x - x_E$ and $\phi_3 = y - y_E$, where (x_E, y_E) is the center of mass of E. Using integration by parts and the aforementioned properties, we obtain:

$$\int_E c^2 \nabla v_h \cdot \nabla \phi_i \, dV = \int_{\partial E} (c^2 \nabla \phi_i) \cdot \mathbf{n} \, v_h \, dS = \sum_{e_k \in \partial E} (c^2 \nabla \phi_i) \cdot \mathbf{n}_k \int_{e_k} v_h \, dS$$

$$= \sum_{e_k \in \partial E} (c^2 \nabla \phi_i) \cdot \mathbf{n}_k |e_k| \frac{v_{a_k} + v_{a_{k+1}}}{2}.$$

The right-hand side of this formula is a linear functional of vector V_E, i.e.

$$\sum_{e_k \in \partial E} (c^2 \nabla \phi_i) \cdot \mathbf{n}_k \, |e_k| \frac{v_{a_k} + v_{a_{k+1}}}{2} = V_E^T \, R_i,$$

with the easily computable vector R_i. Let U_i be the discrete representation of ϕ_i, i.e. values of this linear function at vertices a_k. Combining the last three formulas and using the fact that vector V_E is arbitrary yelds

$$A_E \, U_i = R_i, \quad i = 1, 2, 3. \tag{5}$$

Thus, we have three equations to find a 4×4 symmetric positive definite stiffness matrix. A one-parameter family of solutions (see [19] for more details) is

$$A_E = \frac{1}{c^2 |E|} \sum_{i=2}^{3} R_i \, R_i^T + \frac{1}{c^2 |E|} \zeta P, \quad \zeta > 0,$$

where P is the orthogonal projector on the null space of U_i^T, i.e. $U_i^T P = 0$. In the MFD theory, the energy bounds on the unknown functions v_h are translated to spectral bounds for parameter ζ. The discretization error does depend on ζ, however, any fixed (mesh-independent) choice for ζ leads to an asymptotically second-order spatial discretization scheme. Note that ζ can vary from element to element.

Remark 1. On a cubic element E, the single parameter ζ becomes by a 4×4 matrix with ten parameters [19] and the stabilization term $\zeta \, P$ is replaced by a term $\hat{P} \, Z \, \hat{P}^T$.

On a rectangular element E with edges Δx and Δy, the stiffness matrix A_E has the following form:

$$A_E = \begin{bmatrix} d_0 & -d_1 & -d_2 & -d_3 \\ -d_1 & d_0 & -d_3 & -d_2 \\ -d_2 & -d_3 & d_0 & -d_1 \\ -d_3 & -d_2 & -d_1 & d_0 \end{bmatrix}, \tag{6}$$

where $d_0 = d_1 + d_2 + d_3$ and

$$d_1 = \frac{1}{4}\Big(\frac{\triangle x}{\triangle y} - \frac{\triangle y}{\triangle x} + \zeta\Big), \quad d_2 = \frac{1}{4}\Big(\frac{\triangle x}{\triangle y} + \frac{\triangle y}{\triangle x} - \zeta\Big), \quad d_3 = \frac{1}{4}\Big(-\frac{\triangle x}{\triangle y} + \frac{\triangle y}{\triangle x} + \zeta\Big).$$

2.2 *Elemental Mass Matrix*

To derive an elemental mass matrix M_E, we must obtain algebraic equations similar to (5). The starting point for analysis is to find requirements when the approximate formula

$$V_E^T\, M_E\, U_E \approx \int_E v_h\, u_h\, dV$$

becomes the identity. Following the work of [20], is it sufficient to achieve the identity only when $u_h = 1$. To deal with the volume integral, we must impose additional restriction on v_h. More precisely, it should be integrated exactly with a linearity preserving quadrature whose quadrature points are at the vertices a_k of E. Since the space of functions v_h that satisfy the restrictions formulated above is infinite, one additional restriction is not a problem. Such a quadrature does exists for a general quadrilateral and gives

$$V_E^T\, M_E\, U_1 = |E| \sum_{k=1}^{4} q_k\, v_{a_k},$$

where q_k are the normalized weights. Let Q_1 be the vector collecting q_k. Since V_E is arbitrary, we obtain

$$M_E\, U_1 = Q_1. \tag{7}$$

This is the only equation for the unknown 4×4 matrix M_E. A family of symmetric positive definite solutions is parameterized with a 3×3 symmetric positive definite matrix Z:

$$M_E = \frac{|E|}{4} U_1\, U_1^T + |E| \widetilde{P}\, Z\, \widetilde{P}^T$$

where $\widetilde{P}$ is a 4×3 matrix with linearly independent orthonormal columns orthogonal to U_1, so that $\widetilde{P}^T\, U_1 = 0$.

On a rectangular element, restrictions imposed on the quadrature give

$$U_1^T\, M_E\, U_1 = U_1^T\, Q_1 = |E| \quad \text{and} \quad U_2^T\, M_E\, U_1 = U_2^T\, Q_1 = U_3^T\, M_E\, U_1 = U_3^T\, Q_1 = 0. \tag{8}$$

A symmetric mass matrix that is invariant under rotation of the element by 180° and satisfies (7) for the weights (8) has the following block form

$$M_E = \frac{|E|}{4}\left[\begin{array}{cc|cc} m_1 & m_2 & m_3 & m_4 \\ m_2 & m_5 & m_4 & m_6 \\ \hline m_3 & m_4 & m_1 & m_2 \\ m_4 & m_6 & m_2 & m_5 \end{array}\right], \qquad m_1 + 2m_2 + m_3 + 2m_4 + m_5 + m_6 = 2, \quad (9)$$

where we assumed a clockwise enumeration of vertices starting from the lower left corner. The condition on the sum of the elements is the first condition in (8). The second and the third conditions in (8) can be verified directly by calculating U_i and the matrix-vector products.

3 M-Adaptation

M-adaptation is the process of optimization of the MFD scheme via a special choice of the parameters $m_1, \ldots, m_6$ and ζ. The selection we make reduces the numerical dispersion and/or numerical anisotropy.

3.1 Dispersion and Stability Analysis on a Rectangular Mesh

First, we write the dispersion relationship in a form convenient for analysis. Second, we expand the error in the numerical velocity, $c_h(\boldsymbol{\kappa}) - c$, in powers of κh, where κ is the wave number. Finally, we choose our parameters such that to eliminate the leading terms in this expansion.

Let us consider a rectangular mesh with mesh steps $\triangle x$ and $\triangle y$ and aspect ratio $r = \frac{\triangle y}{\triangle x}$. The second-order time integration scheme is

$$u^{k+1} = u^k + (u^k - u^{k-1}) - c^2 \triangle t^2 \left(D^{-1} M D^{-1}\right) A u^k, \quad (10)$$

where u^k is the global vector of nodal degrees of freedom. Consider a planar wave

$$u(x, y; t) = e^{i(\omega_h t + \kappa_1 x + \kappa_2 y)}, \quad (11)$$

where the numerical frequency ω_h does not have to match the physical frequency ω, i.e. in general $\omega_h \neq \omega = c\kappa$, $\kappa := \sqrt{\kappa_1^2 + \kappa_2^2}$. Substituting nodal values of (11) into Eq. (10), we obtain the conventional dispersion relation. After some algebraic manipulations, it can be written in the following form:

$$2(1 - \cos(\omega_h \triangle t)) = c^2 \triangle t^2 (\mathbf{v}^* D_E^{-1} M_E D_E^{-1} \mathbf{v})(\mathbf{v}^* A_E \mathbf{v}), \ \mathbf{v} = \begin{bmatrix} 1 \\ e^{i\kappa_2 \triangle y} \\ e^{-i\kappa_1 \triangle x} e^{-i\kappa_2 \triangle y} \\ e^{i\kappa_1 \triangle x} \end{bmatrix} \quad (12)$$

and $\mathbf{v}^*$ is the complex conjugate transform of $\mathbf{v}$.

The stability analysis is similar to the dispersion analysis. A bound is imposed on the time step $\triangle t$ to guaranty real values of the wave frequency ω_h. This holds if $\cos(\omega_h \triangle t)$ takes values in the interval $[-1, 1]$:

$$0 \leq c^2 \triangle t^2 (\mathbf{v}^* D_E^{-1} M_E D_E^{-1} \mathbf{v})(\mathbf{v}^* A_E \mathbf{v}) \leq 4 \tag{13}$$

for all $\boldsymbol{\kappa} = (\kappa_1, \kappa_2)$. Since A_E is symmetric semi-positive definite, the product term $\mathbf{v}^* A_E \mathbf{v}$ is real and non-negative. Therefore, the lower bound in (13) is satisfied if M_E is positive definite. Using the norm inequality $\|A\mathbf{v}\| \leq \|A\| \, \|\mathbf{v}\|$, sufficient condition for the upper bound becomes

$$c^2 \triangle t^2 \leq 4 \|\mathbf{v}\|^{-2} \|D_E^{-1} M_E D_E^{-1}\|^{-1} \cdot \|\mathbf{v}\|^{-2} \|A_E\|^{-1}. \tag{14}$$

Here, the operator norm $\|A_E\|$ is subordinate to the standard l^2-vector norm $\|\mathbf{v}\|$. Since $\|\mathbf{v}\|^2 = 4$, we obtain the following stability condition:

$$\nu^2 = \frac{c^2 \triangle t^2}{h^2} \leq \frac{4^3}{h^2} \|D_E^{-1} M_E D_E^{-1}\|^{-1} \|A_E\|^{-1} = 4h^2 \|M_E\|^{-1} \|A_E\|^{-1}. \tag{15}$$

3.2 *Eliminating Dispersion and Anisotropy in the Low Frequency Limit*

Here, we use the dispersion relation (12) to expand the normalized numerical velocity c_h/c, $(c_h = \omega_h/\kappa)$ in powers of κh in the limit $\kappa h \to 0$. For this, we rewrite the $\cos(\cdot)$ as a function of the numerical velocity:

$$\cos(\omega_h \triangle t) = \cos(c_h \kappa \triangle t) = \cos\left(\frac{c_h}{c} \frac{c \triangle t}{h} \kappa h\right) = \cos\left(\frac{c_h}{c} \nu \kappa h\right). \tag{16}$$

Using the derived expressions for the mass and stiffness matrices, we expand the left and right sides of (12) in powers of κh. The power series expansion for the numerical velocity c_h that allows to match the coefficients with the same powers of κh in the previous expansion is computed using Mathematica:

$$\begin{aligned} \frac{c_h(\boldsymbol{\kappa})}{c} =& 1 + \left\{ C_1 \left(\frac{\kappa_1}{\kappa}\right)^2 + C_2 \left(\frac{\kappa_1}{\kappa}\right)\left(\frac{\kappa_2}{\kappa}\right) + C_3 \left(\frac{\kappa_2}{\kappa}\right)^2 + C_4 \left(\frac{\kappa_1}{\kappa}\right)^2 \left(\frac{\kappa_2}{\kappa}\right)^2 \right\} (\kappa h)^2 \\ &+ O\left((\kappa h)^4\right), \end{aligned} \tag{17}$$

where κ_1, κ_2 are components of vector $\bar{\kappa}$ and

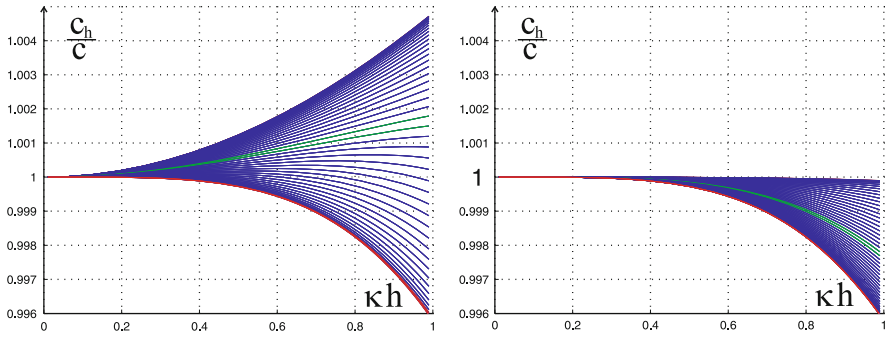


Fig. 1 The dispersion curves for the Courant number $\nu = 0.4$ and $r = 0.5$ for various directions of the wave in the optimized scheme of Guddati and Yue (*left*) and the optimized MFD scheme with $m_1 = m_5 = 1$ (*right*)

$$
\begin{aligned}
C_1 &= \frac{\nu^2 + (-7 + 3m_1 + 6m_2 + 3m_5)}{24}, \\
C_2 &= -\frac{(-2 + m_1 + 2m_2 + 2m_3 + 2m_4 + m_5)r}{4}, \\
C_3 &= \frac{\nu^2 + (-7 + 3m_1 + 6m_4 + 3m_5)r^2}{24}, \quad C_4 = -\frac{2 + 2r^2 - 3r\zeta}{24}.
\end{aligned}
\tag{18}
$$

To eliminate the numerical dispersion at the order $(\kappa h)^2$ in (17) we have to find parameters such that $C_1 = C_2 = C_3 = C_4 = 0$. This is a linear system of four equations in six unknowns:

$$
\begin{aligned}
&m_1 + 2m_2 + m_5 = \tfrac{7}{3} - \tfrac{\nu^2}{3}, \qquad m_1 + 2m_2 + 2m_3 + 2m_4 + m_5 = 2, \\
&m_1 + 2m_4 + m_5 = \tfrac{7}{3} - \tfrac{\nu^2}{3r^2}, \qquad \zeta = \tfrac{2}{3}\left(\tfrac{1}{r} + r\right).
\end{aligned}
\tag{19}
$$

The solution of the above system is defined by two free parameters m_1 and m_5:

$$
\begin{aligned}
&m_2 = \frac{7 - \nu^2}{6} - \frac{1}{2}m_1 - \frac{1}{2}m_5, \qquad m_4 = \frac{7 - \nu^2 r^{-2}}{6} - \frac{1}{2}m_1 - \frac{1}{2}m_5, \\
&m_3 = m_6 = \frac{-8 + \nu^2(1 + r^{-2})}{6} + \frac{1}{2}m_1 + \frac{1}{2}m_5.
\end{aligned}
\tag{20}
$$

Optimization of the MFD scheme with respect to the free parameters will be a focus of the future research.

The conditions (20) are necessary and sufficient for the MFD scheme to eliminate the second-order dispersion. Note that for a strictly rectangular mesh ($r \neq 1$), we obtain $m_2 \neq m_4$; while in the method of Guddati and Yue [17, 21], $m_2 = m_4$. Therefore, the later method is only second-order accurate on rectangular meshes. This is illustrated in Fig. 1. The fan of dispersion curves for the MFD scheme is more compact and close to one for larger value of κh.

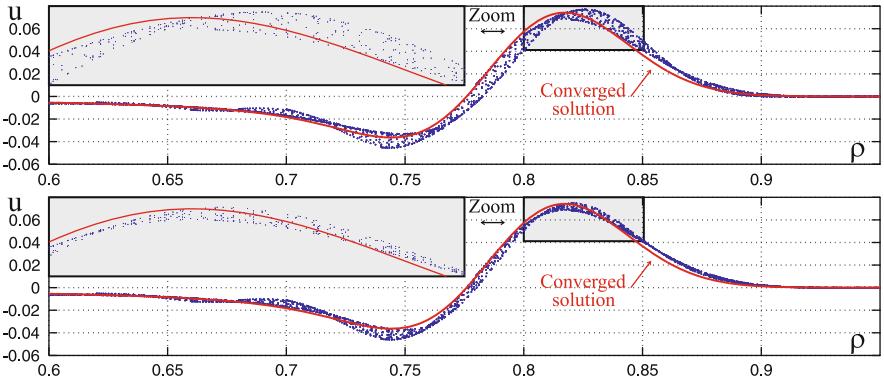


Fig. 2 Displacement as a function of the distance ρ from the origin in the optimized scheme of Guddati and Yue (*top*) and the optimized MFD scheme with parameters $m_1 = m_5 = 1$ (*bottom*) at time $T = 0.8$

4 Numerical Experiments

In the experiment below, we compare the numerical dispersion and the numerical anisotropy for the developed optimized MFD scheme with the scheme proposed in [17, 21].

We consider a problem with radially symmetric initial Gaussian displacement, whose solution is an outward traveling radially symmetric wave. There are several reasons for selecting a Gaussian as part of a solution. For one, it contains a whole range of wavelengths. The mean wavelength λ_{mean} for a Gaussian with a variance σ^2 is analytically computed based on the Fourier transform as $\lambda_{\text{mean}} = \sigma/2\pi$. For two, it contains regions where the values are effectively zero, so it is easy to detect an error propagating into these regions.

The computational domain $[-1, 1]^2$ is partitioned into rectangular elements with $\triangle x = h = 0.02$ and $\triangle y = rh = 0.01$. The initial displacement is chosen to be Gaussian and the boundary conditions are zero:

$$\begin{cases} u(\mathbf{x}; t = 0) = \text{Gauss}(|\mathbf{x}|, \sigma, 0), \\ u_t(\mathbf{x}; t = 0) = 0, \end{cases} \qquad \sigma = \frac{\lambda_{\text{mean}}}{2\pi} = \frac{N_\lambda h}{2\pi},$$

where $N_\lambda = 10$ is the number of points per mean wavelength. The discretization time step $\triangle t = 0.008$ corresponds to the Courant number $\nu = 0.4$.

The numerical experiment continues until $T = 0.8$. In this time the wave travels a distance of $\frac{T_c}{\lambda} = 4$ mean wavelengths. Instead of plotting a three dimensional figure (where the error is difficult to see), on Fig. 2 we plot the displacement u_h as a function of the distance from the origin $\rho = |\mathbf{x}|$. Due to the numerical anisotropy, the dependence $u_h(\rho)$ does not represent a line but a band. The thickness of the band is a measure of the anisotropy and is smaller in the MFD scheme.

5 Conclusions

We developed a new optimization technique, dubbed m-adaptation, for the numerical solution of the acoustic wave equation on rectangular meshes. This technique is based on selection of an optimal scheme in the rich family of MFD schemes. This scheme is second-order accurate in space and time and has the fourth-order numerical dispersion. We found a two-parameter subfamily of schemes with such properties. The optimization of this subfamily will be a topic of a future research.

Acknowledgements This and previous work was carried out under the auspices of the National Nuclear Security Administration of the U.S. Department of Energy at Los Alamos National Laboratory under Contract No. DE-AC52-06NA25396 and the DOE Office of Science Advanced Scientific Computing Research (ASCR) Program in Applied Mathematics Research.

References

1. O. Holberg, "Computational aspects of the choice of operator and sampling interval for numerical differenciation in large-scale simulation of wave phenomena," *Geophys. Prospect.*, vol. 35, pp. 629–655, 1987.
2. F. Hu, M. Hussaini, and J. L. Manthey, "Low-dissipation and low-dispersion Runge-Kutta schemes for Computational Acoustics," *J. Comput. Phys.*, vol. 124, pp. 177–191, 1996.
3. G. Ashcroft and X. Zhang, "Optimized prefactored compact schemes," *J. Comput. Phys.*, vol. 190, pp. 459–477, 2003.
4. C. Bogey and C. Bailly, "A family of low dispersive and low dissipative explicit schemes for flow and noise computation," *J. Comput. Phys.*, vol. 194, pp. 194–214, 2004.
5. S. Lele, "Compact finite difference schemes with spectral-like resolution," *J. Comput. Phys.*, vol. 103, pp. 16–42, 1992.
6. C.-H. Jo, C. Shin, and J. H. Suh, "An optimal 9-point, finite difference, frequency-space, 2-D scalar wave extrapolator," *Geophysics*, vol. 61, pp. 529–537, 1996.
7. I. Stekl and R. Pratt, "Accurate viscoelastic modeling by frequency-domain finite difference using rotated oprators," *Geophysics*, vol. 63, no. 5, pp. 1779–1794, 1998.
8. A. Sescu, R. Hixon, and A. A. Afjeh, "Multidimensional optimization of finite difference schemes for Computational Aeroacoustics," *J. Comput. Phys.*, pp. 4563–4588, 2008.
9. R. Mullen and T. Belytschko, "Dispersion analysis of Finite Element semidiscretizations of the two-dimensional wave equation," *Int. J. Numer. Meth. Eng.*, vol. 18, pp. 11–29, 1982.
10. M. A. Christon and T. E. Voth, "Results of von Neumann analyses for reproducing kernel semi-discretizations," *Int. J. Numer. Meth. Eng.*, vol. 47, pp. 1285–1301, 2000.
11. D. Bueche, N. Sukumar, and B. Moran, "Dispersive properties of the natural element method," *Computational Mechanics*, vol. 25, pp. 207–219, 2000.
12. G. Seriani, E. Priolo, and A. Pregarz, "Modelling waves in anisotropic media by a spectral element method," in *Mathematical and numerical aspects of wave propagation (Mandelieu-La Napoule, 1995)*, (Philadelphia, PA), pp. 289–298, SIAM, 1995.
13. D. Komatitsch, R. Martin, J. Tromp, M. Taylor, and B. Wingate, "Wave propagation in 2-D elastic media using a spectral element method with triangles and quadrangles," *J. Comput. Acoust.*, vol. 9, no. 2, pp. 703–718, 2001.
14. C. Farhat, I. Harari, and L. P. Franca, "The discontinuous enrichment method," *Comput. Methods Appl. Mech. Engrg.*, vol. 190, no. 48, pp. 6455–6479, 2001.

15. C. J. Gittelson, R. Hiptmair, and I. Perugia, "Plane wave discontinuous Galerkin methods: analysis of the *h*-version," *M2AN Math. Model. Numer. Anal.*, vol. 43, no. 2, pp. 297–331, 2009.
16. S. Krenk, "Dispersion-corrected explicit integration of the wave equation," *Comput. Methods Appl. Mech. Engrg.*, vol. 191, pp. 975–987, 2001.
17. B. Yue and M. N. Guddati, "Dispersion-reducing finite elements for transient acoustics," *J. Acoust. Soc. Am.*, vol. 118, pp. 2132–2141, 2005.
18. V. Gyrya and K. Lipnikov, "M-adaptation method for acoustic wave equation on square meshes," *Report LA-UR 12-10047, submitter to J. Acoust. Soc. Am.*, 2012.
19. F. Brezzi, A. Buffa, and K. Lipnikov, "Mimetic finite differences for elliptic problems," *M2AN Math. Model. Numer. Anal.*, vol. 43, no. 2, pp. 277–295, 2009.
20. G. Cohen, P. Joly, J. E. Roberts, and N. Tordjman, "Higher order triangular finite elements with mass lumping for the wave equation," *SIAM J. Numer. Anal.*, vol. 38, no. 6, pp. 2047–2078 (electronic), 2001.
21. M. N. Guddati and B. Yue, "Modified integration rules for reducing dispersion error in finite element methods," *Comput. Methods Appl. Mech. Engrg.*, vol. 193, pp. 275–287, 2004.

Applications of Nonvariational Finite Element Methods to Monge–Ampère Type Equations

T. Pryer

Abstract The goal of this work is to illustrate the application of the nonvariational finite element method to a specific Monge–Ampère type nonlinear partial differential equation. The equation we consider is that of prescribed Gauss curvature however the method can be generalised to any Monge–Ampère operator.

1 Introduction and Problem Setting

The *nonvariational finite element method* (NVFEM) introduced in [10] is a numerical method designed for problems posed over an open, bounded domain $\Omega \subset \mathbb{R}^d$ of the form

$$\mathbf{A}(\mathbf{x}):\mathrm{D}^2 u(\mathbf{x}) = f(\mathbf{x}) \tag{1}$$

where $\mathrm{D}^2 u$ denotes the Hessian of the function u and for each $\mathbf{x} \in \Omega$, $\mathbf{A}(\mathbf{x}) \in \mathbb{R}^{d\times d}$ is symmetric positive definite matrix. The operation $\mathbf{B}:\mathbf{C} = \mathrm{trace}(\mathbf{B}^\intercal\mathbf{C})$ is the Frobenius inner product between two $d \times d$ matrices. Classical finite element methods are applicable to this problem if we assume the coefficient matrix $\mathbf{A}$ is differentiable. In this case we may rewrite (1) in *variational* or *divergence* form via the introduction of an advection term since

$$f(\mathbf{x}) = \mathbf{A}(\mathbf{x}):\mathrm{D}^2 u(\mathbf{x}) = \mathrm{div}\big(\mathbf{A}(\mathbf{x})\nabla u(\mathbf{x})\big) - \mathrm{D}\mathbf{A}(\mathbf{x})\nabla u(\mathbf{x}) \tag{2}$$

where

$$\mathrm{D}\mathbf{A}(\mathbf{x}) = \left(\sum_{i=1}^{d} \partial_i a_{i,1}(\mathbf{x}), \ldots, \sum_{i=1}^{d} \partial_i a_{i,d}(\mathbf{x})\right). \tag{3}$$

T. Pryer (✉)
University of Kent, Canterbury, UK
e-mail: T.Pryer@kent.ac.uk

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_47,

Note that we are using the convention that $\nabla\phi = \left(\partial_1\phi, \ldots, \partial_d\phi\right)^\intercal$ is the column vector formed of first order partial derivatives of a d–multivariate function ϕ.

The introduction of the advection term may result in the variational problem becoming advection dominated. This is undesirable in the finite element context and stabilisation terms become necessary to derive a viable numerical method [7, cf.]. Interestingly if $\|\mathrm{D}\mathbf{A}\|_{\mathrm{L}_\infty(\Omega)} \gg \|\mathbf{A}\|_{\mathrm{L}_\infty(\Omega)}$ applying the NVFEM to (1) does not result in an unstable scheme, whereas applying a standard FEM to (2) does. This is numerically demonstrated in [10, § 4.2]. It may even be the case that $\mathbf{A}$ is not differentiable, in which case the standard FEM cannot be applied.

The fully nonlinear problem

$$\mathscr{F}(\mathrm{D}^2 u) = 0 \tag{4}$$

is related to the nonvariational problem (1) [6]. This can be seen by applying a Newton linearisation to (4), resulting in a sequence of linear nonvariational PDEs:

$$\mathscr{F}'(\mathrm{D}^2 u^n) : \mathrm{D}^2\left(u^{n+1} - u^n\right) = -\mathscr{F}(\mathrm{D}^2 u^n), \tag{5}$$

where $\mathscr{F}'$ is the Fréchet derivative of $\mathscr{F}$.

The Monge–Ampère operators are an extremely interesting class of fully nonlinear PDE. These arise from differential geometry and optimal transport problems; they take the form

$$\mathscr{F}(\mathrm{D}^2 u, \nabla u, u, \mathbf{x}) := \det\left(\mathrm{D}^2 u\right) - f(\nabla u, u, \mathbf{x}) = 0. \tag{6}$$

For example the Monge–Ampère–Dirichlet (MAD) problem is the case when $f = f(\mathbf{x})$ and (6) is coupled with a Dirichlet type boundary condition ($u = g$ on $\partial\Omega$). This particular equation is a prototypical example of a fully nonlinear PDE.

There are a variety of numerical methods available for the more general Monge–Ampère class of fully nonlinear PDE (6). In [13] the author proposes a wide stencil finite difference scheme. In [3] a C^1 finite element scheme based on the Argyris element is used. In [8,9] the authors construct numerical approximations of solutions to sequences of quasilinear biharmonic equations. This is very reminiscent of the vanishing viscosity method first studied for use in fully nonlinear first order PDEs. The method is aptly named the vanishing moment method. More recently in [4] a consistent penalisation method has been introduced for these problems. Finally, in [2] the author uses a *Laplacian relaxation* technique to study these equations.

For the Monge–Ampère type equation (6) to be well posed we require $\Omega \subset \mathbb{R}^d$ to be a convex domain and $f > 0$. The Monge–Ampère operator is elliptic over the cone of strictly convex functions in Ω and under the constraints above will admit a unique convex viscosity solution [6].

In this work we will study the equation of prescribed Gauss curvature. This arises from the problem of finding a function u such that the graph of u has a specified Gaussian curvature K. In this case we have that

$$K = \frac{\det \mathrm{D}^2 u}{\left(1 + |\nabla u|^2\right)^{(d+2)/2}} \tag{7}$$

and hence

$$\mathscr{F}(\mathrm{D}^2 u, \nabla u, \mathbf{x}) := \det \mathrm{D}^2 u - K\left(1 + |\nabla u|^2\right)^{(d+2)/2}. \tag{8}$$

Note that $K = K(u, \mathbf{x})$.

The linearisation of this problem can be calculated in a direction v as

$$\begin{aligned} \mathscr{F}'(\mathrm{D}^2 u, \nabla u, \mathbf{x}) : \mathrm{D}^2 v &= \lim_{\epsilon \to 0} \frac{1}{\epsilon}\Big(\mathscr{F}(\mathrm{D}^2 u + \epsilon \mathrm{D}^2 v, \nabla u + \epsilon \nabla v, \mathbf{x}) - \mathscr{F}(\mathrm{D}^2 u, \nabla u, \mathbf{x})\Big) \\ &= \operatorname{Cof} \mathrm{D}^2 u : \mathrm{D}^2 v + (d+2)\, K\left(\left(1 + |\nabla u|^2\right)^{d/2} (\nabla u)^{\intercal} \nabla v\right) \end{aligned} \tag{9}$$

and thus the linearisation is elliptic if $\operatorname{Cof} \mathrm{D}^2 u$ is an elliptic operator. This holds for convex u.

2 Discretisation

The process of discretisation can be sought in two ways. We may look at the continuous problem and discretise this directly, resulting in a system of nonlinear equations, or we may first linearise the problem and discretise from there. Discretising the nonlinear problem directly is certainly possible but is more technical, as discussed in [4]. For brevity we will perform a Newton linearisation on (6) and discretise the sequence of linear nonvariational PDEs in a similar light to [11].

Let $\mathscr{T}$ be a conforming, shape regular triangulation of Ω, namely, $\mathscr{T}$ is a finite family of sets such that

1. $K \in \mathscr{T}$ implies K is an open simplex (segment for $d = 1$, triangle for $d = 2$, tetrahedron for $d = 3$),
2. For any $K, J \in \mathscr{T}$ we have that $\overline{K} \cap \overline{J}$ is a full sub-simplex (i.e., it is either $\emptyset$, a vertex, an edge, a face, or the whole of $\overline{K}$ and $\overline{J}$) of both $\overline{K}$ and $\overline{J}$ and
3. $\bigcup_{K \in \mathscr{T}} \overline{K} = \overline{\Omega}$.

We use the convention where $h : \Omega \to \mathbb{R}$ denotes the *meshsize function* of $\mathscr{T}$, i.e.,

$$h(\mathbf{x}) := \max_{\overline{K} \ni \mathbf{x}} h_K, \tag{10}$$

where h_K is the maximal diameter of an element K.

Definition 1 (FE spaces). Let $\mathbb{P}^p(\mathscr{T})$ denote the space of piecewise polynomials of degree k over the triangulation $\mathscr{T}$ of Ω. We introduce the *finite element spaces*

$$\mathbb{V} = \mathbb{P}^p(\mathscr{T}) \cap \mathrm{C}^0(\Omega) \cap \mathrm{H}_0^1(\Omega) \text{ and } \mathbb{W} = \mathbb{P}^p(\mathscr{T}) \cap \mathrm{C}^0(\Omega) \tag{11}$$

to be the usual space of continuous piecewise polynomial functions and

$$\mathbb{S} := \mathbb{V} \times \mathbb{W}^{d\times d}. \tag{12}$$

Remark 1 (generalised Hessian). Given a function $v \in \mathrm{H}^2(\Omega)$, let $\mathbf{n} : \partial\Omega \to \mathbb{R}^d$ be the outward pointing normal of Ω then the Hessian of v, $\mathrm{D}^2 v$, satisfies the following identity:

$$\left\langle \mathrm{D}^2 v \,|\, \phi \right\rangle = -\int_\Omega \nabla v \otimes \nabla \phi + \int_{\partial\Omega} \nabla v \otimes \mathbf{n}\, \phi \quad \forall\, \phi \in \mathrm{H}^1(\Omega) \tag{13}$$

where $\langle \cdot \,|\, \cdot \rangle$ denotes an appropriate duality pairing. It follows that we may weaken the regularity assumptions to $v \in \mathrm{H}^1(\Omega) \cap \mathrm{H}^1(\partial\Omega)$.

Definition 2 (finite element Hessian). From Remark 1 and in view of Riesz representation theorem we may define for a given $V \in \mathbb{V}$, the *finite element Hessian*, $\mathbf{H}[V] \in \mathbb{W}^{d\times d}$, such that

$$\int_\Omega \mathbf{H}[V]\Phi = \left\langle \mathrm{D}^2 V \,|\, \Phi \right\rangle \quad \forall\, \Phi \in \mathbb{W}. \tag{14}$$

Proposition 1 (symmetry of the finite element Hessian). *The finite element Hessian is symmetric, that is for each $V \in \mathbb{V}$*

$$\int_\Omega \mathbf{H}[V]\Phi = \int_\Omega \big(\mathbf{H}[V]\big)^{\mathsf{T}}\Phi \quad \forall\, \Phi \in \mathbb{W}. \tag{15}$$

In view of the constraints to the continuous problem (8) to admit a unique solution it is also necessary to construct a discrete notion of convexity. This has been developed in [1] and is naturally passed down from the concept of distributional convexity.

Definition 3 (finite element convexity [1]). A function, $v \in \mathrm{H}^1(\Omega) \cap \mathrm{H}^1(\partial\Omega)$, is said to be *finite element convex* if

$$\int_\Omega \mathbf{H}[v]\Phi \text{ is positive semidefinite} \quad \forall\, \Phi \in \mathbb{W} \tag{16}$$

where $\Phi \geq 0$ on Ω. It is strictly finite element convex if (16) is positive definite.

Definition 2 allows us to construct what is essentially a 2–0 mixed method where the Hessian of the solution to (6) is treated as an auxiliary variable in the formulation (as opposed to the 1–1 mixed methods commonly found in the literature by decoupling a second order PDE into a system of first order PDEs [5, cf.]).

Given the linearisation (9) we formulate the problem in the discrete setting as follows: Given an initial guess $\left(U^0, \mathbf{H}[U^0]\right) \in \mathbb{S}$ that is strictly finite element convex (16), for $n \in \mathbb{N}$ find $\left(U^n, \mathbf{H}[U^n]\right) \in \mathbb{S}$ such that

$$\int_\Omega \mathbf{H}[U^n]\Phi + \int_\Omega \nabla U^n \otimes \nabla \Phi - \int_{\partial\Omega} \left(\nabla U^n\right)^{\mathsf{T}} \mathbf{n}\, \Phi = 0 \quad \forall\, \Phi \in \mathbb{W} \tag{17}$$

$$\int_\Omega \mathscr{F}'\left(\mathbf{H}[U^{n-1}], \nabla U^{n-1}\right) : \mathbf{H}[U^n - U^{n-1}] + \mathscr{F}\left(\mathbf{H}[U^{n-1}], \nabla U^{n-1}\right)\Psi = 0 \quad \forall\, \Psi \in \mathbb{V}. \tag{18}$$

Where for the problem of prescribed Gaussian curvature (18) is

$$\begin{aligned}\int_\Omega \mathscr{F}'\left(\mathbf{H}[U^{n-1}], \nabla U^{n-1}\right) : \mathbf{H}[U^n - U^{n-1}] + \mathscr{F}\left(\mathbf{H}[U^{n-1}], \nabla U^{n-1}\right)\Psi \\ = \int_\Omega \operatorname{Cof} \mathbf{H}[U^{n-1}] : \mathbf{H}[U^n - U^{n-1}]\Psi \\ + \int_\Omega 2dK\left(1 + \left|\nabla U^{n-1}\right|^2\right)^{d/2} \left(\nabla U^{n-1}\right)^{\mathsf{T}} \nabla\left(U^n - U^{n-1}\right)\Psi \\ + \int_\Omega \left(\det \mathbf{H}[U^{n-1}] - K\left(1 + \left|\nabla U^{n-1}\right|^2\right)^{(d+2)/2}\right)\Psi.\end{aligned} \tag{19}$$

Due to the symmetry property given in Proposition 1 we may simplify the problem somewhat to seeking only the upper (or lower) triangular parts of the finite element Hessian. This reduces $\mathbb{S} = \mathbb{V} \times \mathbb{W}^{(d^2+d)/2}$.

Theorem 1 (solvability of the discrete system [14]). *Let $U \in \mathbb{V}$ be the nonvariational finite element approximation to u, the solution of the elliptic nonvariational problem (1). Then we have a discrete inf–sup condition, that is the linear system is always invertible. Hence, assuming the linearisation maintains ellipticity, the discrete problem (17)–(18) is well posed.*

3 Numerical Experiments

In this section we detail numerical experiments on the formulation (17)–(18).

We will consider the case $d = 2$ and when $K > 0$ is some prescribed curvature. In each of the experiments we choose $p = 2$, i.e., $\mathbb{V}$ consists of piecewise quadratic functions. The domain Ω is taken as a square whose size differs on each of the experiments and the triangulation $\mathscr{T}$ is unstructured. All of the numerical experiments have been conducted using the DOLFIN environment of the finite element package FEniCS [12].

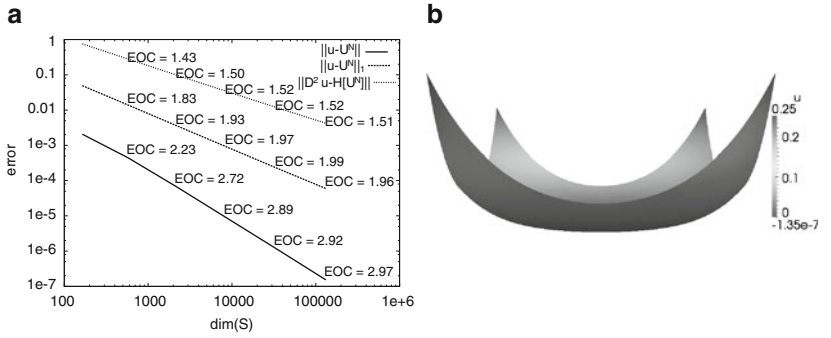


Fig. 1 In this experiment we choose a convex solution u which classically solves the equation of prescribed Gauss curvature (8) over the square $[-.5,.5]^2$. That is, we fix $u = |\mathbf{x}|^4$ and calculate $K = K(\mathbf{x}, u)$. We solve the discrete problem over a sequence of concurrently refined meshes and ascertain the errors and convergence rates for the problem in $\mathrm{L}_2(\Omega)$, $\mathrm{H}_0^1(\Omega)$ and a discrete $\mathrm{H}^2(\Omega)$ seminorm. Notice that $\left\|u - U^N\right\| \approx \mathrm{O}(h^3)$, $\left|u - U^N\right|_1 \approx \mathrm{O}(h^2)$ and $\left\|\mathrm{D}^2 u - \mathbf{H}[U^N]\right\| \approx \mathrm{O}(h^{1.5})$. (**a**) Errors and convergence rates for the problem, $p = 2$. (**b**) Solution plot

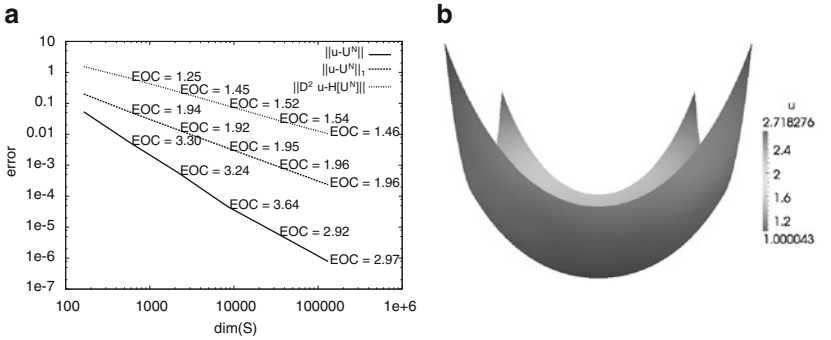


Fig. 2 In this experiment we choose a convex solution u which classically solves the equation of prescribed Gauss curvature (8) over the square $[-.5,.5]^2$. That is, we fix $u = \exp(|\mathbf{x}|^2/2)$ and calculate $K = K(\mathbf{x}, u)$. We solve the discrete problem over a sequence of concurrently refined meshes and ascertain the errors and convergence rates for the problem in $\mathrm{L}_2(\Omega)$, $\mathrm{H}_0^1(\Omega)$ and a discrete $\mathrm{H}^2(\Omega)$ seminorm. Notice that $\left\|u - U^N\right\| \approx \mathrm{O}(h^3)$, $\left|u - U^N\right|_1 \approx \mathrm{O}(h^2)$ and $\left\|\mathrm{D}^2 u - \mathbf{H}[U^N]\right\| \approx \mathrm{O}(h^{1.5})$. (**a**) Errors and convergence rates for the problem, $p = 2$. (**b**) Solution plot

In Figs. 1–2 we construct classical solutions to (8) in order to look at the numerical convergence of the method. In Fig. 3 we consider K as a constant over the domain $[-.57,.57]^2$. These results can then be compared with the two other numerical studies found in the literature on prescribed Gauss curvature [2, 9]. In these experiments the authors note that the problem (8) is well posed only for $K \leq K^{\max}$ and estimate the value of $K^{\max}$ by asserting when the numerical algorithm proposed breaks down.

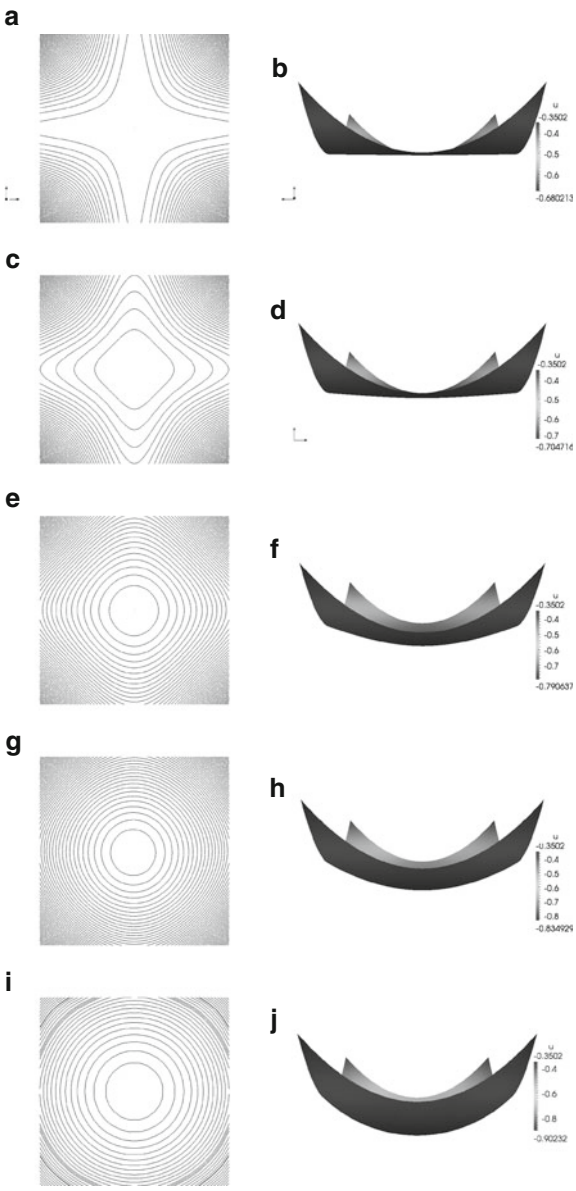


Fig. 3 In this experiment we consider the case K is constant. We take boundary values $g = |\mathbf{x}|^2 - 1$ and fix $h \approx 0.009$. We implement the discrete problem over an unstructured mesh of the square $[-0.57, 0.57]^2$. We choose various values of $K > 0$ and display a contour plot together with a side view of the discrete solution. Note that the numerical algorithm fails to converge for $K = 2$.

(**a**) Contour plot for $K = 0.01$.

(**b**) Solution plot for $K = 0.01$.

(**c**) Contour plot for $K = 0.1$.

(**d**) Solution plot for $K = 0.1$.

(**e**) Contour plot for $K = 0.5$.

(**f**) Solution plot for $K = 0.5$.

(**g**) Contour plot for $K = 1.0$.

(**h**) Solution plot for $K = 1.0$.

(**i**) Contour plot for $K = 1.5$.

(**j**) Solution plot for $K = 1.5$

The initial guess to any Newton iteration is paramount due to the well known *overshoot* property. In the case of Monge–Ampère type linearisations it's especially important since (discrete) convexity must be maintained during the iterative procedure for the problem to remain well posed. In each of the tests below we initialise the algorithm by approximating the solution of the MAD problem over the initial mesh as detailed in [11].

$$\begin{aligned} \det \mathrm{D}^2 u &= K \quad \text{in } \Omega \\ u &= g \quad \text{on } \partial\Omega. \end{aligned} \tag{20}$$

The iterative procedure given by the discrete problem (17)–(18) is terminated when two concurrent iterates satisfy $\left\| U^n - U^{n-1} \right\|_{\mathrm{L}_\infty(\Omega)} \leq 10^{-10}$.

Remark 2 (Comparison of convergence rates to the MAD problem). The results show the error committed by the numerical solution measured in $\mathrm{L}_2(\Omega)$ converge optimally, $\mathrm{O}(h^3)$, for the benchmarked solutions to the prescribed Gauss curvature equation (compare Figs. 1 and 2). For the MAD problem it has been observed that the $\mathrm{L}_2(\Omega)$ error converges sub-optimally (see [11, cf.]).

References

1. Néstor E. Aguilera and Pedro Morin. On convex functions and the finite element method. *SIAM J. Numer. Anal.*, 47(4):3139–3157, 2009.
2. Gerard Awanou. Pseudo time continuation and time marching methods for monge–ampere type equations. *In revision - tech report available on http://www.math.niu.edu/~awanou/*, 2012.
3. Klaus Böhmer. On finite element methods for fully nonlinear elliptic equations of second order. *SIAM J. Numer. Anal.*, 46(3):1212–1249, 2008.
4. Susanne C. Brenner, Thirupathi Gudi, Michael Neilan, and Li-yeng Sung. C^0 penalty methods for the fully nonlinear Monge-Ampère equation. *Math. Comp.*, 80(276):1979–1995, 2011.
5. Franco Brezzi and Michel Fortin. *Mixed and hybrid finite element methods*, volume 15 of *Springer Series in Computational Mathematics*. Springer-Verlag, New York, 1991.
6. Luis A. Caffarelli and Xavier Cabré. *Fully nonlinear elliptic equations*, volume 43 of *American Mathematical Society Colloquium Publications*. American Mathematical Society, Providence, RI, 1995.
7. Alexandre Ern and Jean-Luc Guermond. *Theory and practice of finite elements*, volume 159 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2004.
8. Xiaobing Feng and Michael Neilan. Mixed finite element methods for the fully nonlinear Monge-Ampère equation based on the vanishing moment method. *SIAM J. Numer. Anal.*, 47(2):1226–1250, 2009.
9. Xiaobing Feng and Michael Neilan. Vanishing moment method and moment solutions for fully nonlinear second order partial differential equations. *J. Sci. Comput.*, 38(1):74–98, 2009.
10. Omar Lakkis and Tristan Pryer. A finite element method for second order nonvariational elliptic problems. *SIAM J. Sci. Comput.*, 33(2):786–801, 2011.
11. Omar Lakkis and Tristan Pryer. A nonvariational finite element method for fully nonlinear elliptic problems. *Submitted - tech report available on ArXiV http://arxiv.org/abs/1103.2970*, 2012.
12. Anders Logg and Garth N. Wells. DOLFIN: automated finite element computing. *ACM Trans. Math. Software*, 37(2):Art. 20, 28, 2010.
13. Adam M. Oberman. Wide stencil finite difference schemes for the elliptic Monge-Ampère equation and functions of the eigenvalues of the Hessian. *Discrete Contin. Dyn. Syst. Ser. B*, 10(1):221–238, 2008.
14. Tristan Pryer. Recovery methods for evolution and nonlinear problems. *DPhil Thesis, University of Sussex*, 2010.

Geodesic Finite Elements in Spaces of Zero Curvature

O. Sander

Abstract We investigate geodesic finite elements for functions with values in a space of zero curvature, like a torus or the Möbius strip. Unlike in the general case, a closed-form expression for geodesic finite element functions is then available. This simplifies computations, and allows us to prove optimal estimates for the interpolation error in 1d and 2d. We also show the somewhat surprising result that the discretization by Kirchhoff transformation of the Richards equation proposed in Berninger et al. (SIAM J Numer Anal 49(6):2576–2597, 2011) is a discretization by geodesic finite elements in the manifold $\mathbb{R}$ with a special metric.

1 Geodesic Finite Elements

Let Ω be an open and connected subset of $\mathbb{R}^d$ with a Lipschitz boundary, and let M be a smooth, connected, m-dimensional manifold. For some smooth embedding of M into a Euclidean space $\mathbb{R}^k$ we define the Sobolev spaces

$$H^p(\Omega, M) := \{v \in H^p(\Omega, \mathbb{R}^k) \mid v(x) \in M \text{ a.e.}\},$$

and note that they are independent of the embedding [1]. Note that the $H^p(\Omega, M)$ have the structure of nonlinear manifolds [4].

Let M be equipped with a Riemannian metric, which turns M into a metric space with distance function $\mathrm{dist} : M \times M \to \mathbb{R}$. For the numerical treatment of partial differential equations for functions in $H^1(\Omega, M)$, geodesic finite elements

O. Sander (✉)
Freie Universität Berlin, Arnimallee 6, 14195 Berlin, Germany
e-mail: sander@mi.fu-berlin.de

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_48,

have been introduced in [5]. Let G be a conforming grid of Ω with simplex elements only. Geodesic finite elements are defined in two steps. The crucial first one is a generalization of linear interpolation to functions from simplices to M. Let $\Delta \subset \mathbb{R}^{d+1}$ be the d-dimensional standard simplex.

Definition 1. Let M be a Riemannian manifold and $\operatorname{dist}(\cdot,\cdot) : M \times M \to \mathbb{R}$ a distance function on M. For values $v_1, \dots, v_{d+1} \in M$ we call

$$\Upsilon : \Delta \to M$$

$$\Upsilon(v_1, \dots, v_{d+1}; w) = \operatorname*{arg\,min}_{q \in M} \sum_{i=1}^{d+1} w_i \operatorname{dist}(v_i, q)^2 \tag{1}$$

simplicial geodesic interpolation between the values $v_1, \dots, v_{d+1}$ on M.

The interpolation function Υ is well defined if the corner values $v_1, \dots, v_{d+1}$ are "close together" in a certain sense. A precise statement, which involves the curvature of M, is given in [5].

In a second step, this interpolation scheme is used to construct global finite element spaces.

Definition 2 (Geodesic Finite Elements). Let G be a simplicial grid on Ω, and let M be a Riemannian manifold. We call $v_h : \Omega \to M$ a geodesic finite element function if it is continuous, and for each element $T \in G$ the restriction $v_h|_T$ is a geodesic simplicial interpolation in the sense that

$$v_h|_T(x) = \Upsilon\big(v_{T,1}, \dots, v_{T,d+1}; \mathscr{F}_T(x)\big),$$

where $\mathscr{F}_T : T \to \Delta$ is affine and the $v_{T,i}$ are values in M.

A detailed investigation of these functions is given in [5]. In numerical experiments, optimal discretization error behavior can be observed.

A disadvantage of the geodesic finite element method is the implicit definition of the interpolation functions (1). This makes their handling challenging both for theoretical investigations and for practical computations. In this article we study the special case that M is a manifold of zero curvature. In this case, a closed-form expression for the interpolation function Υ is available. Consequently, the handling of geodesic finite element functions is simplified considerably. It also follows that certain nonlinear scaling techniques for PDEs can be interpreted as geodesic finite elements.

Chapter 2 introduces spaces of zero curvature and derives the closed-form expression for geodesic interpolation. In Chap. 3 we prove optimal interpolation error bounds in various Sobolev norms if $d \leq 2$. In Chap. 4, finally, we give a reinterpretation of a discretization of the Richards equation based on Kirchhoff transformation introduced in [2]. We show that the discretization there is actually a discretization by geodesic finite elements with $M = \mathbb{R}$ and a special metric.

2 Spaces of Zero Curvature

Let M be a Riemannian manifold. We say that M has curvature zero if all sectional curvatures[1] at a point x are zero for all $x \in M$. Examples of such manifolds are:

1. All one-dimensional manifolds.
2. The complete and connected spaces of zero curvature can be classified precisely:

   Theorem 1 (Wolf [7], Theorems 2.4.9, 3.1.3). *Let M be a Riemannian manifold of dimension m. Then M is complete, connected and of constant curvature zero if and only if it is isometric to a quotient $\mathbb{R}^m/\Gamma$, where Γ is a discrete subgroup of the group of isometries of $\mathbb{R}^m$, and acting without fixed points.*

   This includes the tori, the Möbius strip and the Klein bottle.
3. Let W be a connected open subset of $\mathbb{R}^m$, and let

$$\phi : W \to \mathbb{R}^m$$

   be C^∞ and such that the tangent map $\nabla\phi(x) : T_x W \to T_{\phi(x)}\mathbb{R}^m$ is invertible for all $x \in W$. At any $x \in W$ define a metric

$$g_x(v, w) = v^T (\nabla\phi(x))^T \nabla\phi(x) w \qquad \text{for all } v, w \in T_x W.$$

   By the assumptions on ϕ the bilinear form g is indeed a metric. As a special case of Theorem 2 below, the manifold (W, g) has zero curvature.

The characterizing property of zero-curvature spaces is that they are locally isometric to Euclidean space. This is formalized by the following theorem, which is a special case of Theorem 2.4.11 from [7].

Theorem 2. *Let M be a Riemannian manifold. Then M is of constant curvature zero if and only if for each $x \in M$ there are local coordinates on a neighborhood of x in which the metric is given by the identity matrix. These coordinate functions are isometries.*

With this result we can derive an explicit formula for geodesic simplicial interpolation in manifolds of zero curvature.

Lemma 1. *Let M be an m-dimensional Riemannian manifold of zero curvature, and let $W \subset M$ be open and such that there exists an isometry $\phi : W \to \mathbb{R}^m$. Let $v_1, \ldots, v_{d+1} \in W$. Then*

$$\Upsilon(v_1, \ldots, v_{d+1}; w) = \phi^{-1}\Big(\sum_{i=1}^{d+1} w_i \phi(v_i)\Big). \tag{2}$$

[1] See [7] or any standard textbook on differential geometry for a definition.

Proof. The inverse ϕ^{-1} exists because ϕ is an isometry [7]. We use the following short result. Let P and U be two sets, $h : P \to U$ a bijection, and $H : U \to \mathbb{R}$. Then

$$u^* = \operatorname*{arg\,min}_{u \in U} H(u)$$

is equivalent to

$$p^* = h^{-1}(u^*) = h^{-1}(\operatorname*{arg\,min}_{u \in U} H(u)) = \operatorname*{arg\,min}_{p \in P} H(h(p)).$$

Setting $P = W$ and $U = \phi(W) \subset \mathbb{R}^m$ we can now simply compute

$$\begin{aligned}
\Upsilon(v_1, \ldots, v_{d+1}; w) &= \operatorname*{arg\,min}_{p \in M} \sum_{i=1}^{d+1} w_i \operatorname{dist}(v_i, p)^2 \\
&= \operatorname*{arg\,min}_{p \in M} \sum_{i=1}^{d+1} w_i |\phi(p_i) - \phi(p)|^2 \\
&= \phi^{-1}\Big(\operatorname*{arg\,min}_{u \in \mathbb{R}^m} \sum_{i=1}^{d+1} w_i |\phi(p_i) - u|^2 \Big),
\end{aligned}$$

for all $w \in \Delta$. The last minimization problem is nothing but linear interpolation in $\mathbb{R}^m$, and the assertion is shown.

Hence if an isometric coordinate function ϕ is known then the minimization problem (1) can be replaced by the much simpler formula (2). The computation of derivatives of Υ simplifies correspondingly. In particular, the indirect method of computing derivatives through the implicit function theorem [5, Chap. 5] becomes unnecessary. We note that in applications, suitable isometries are frequently available (see, e.g., Chap. 4).

3 Interpolation Error

In this chapter we prove optimal bounds for the interpolation error. We restrict our attention to one- and two-dimensional domains, in order to be able to work with the standard interpolation operator.

Let $d \in \{1, 2\}$. By the Sobolev embedding theorems functions in $H^2(\Omega, M)$ are continuous and we can define the interpolation operator

$$\begin{aligned}
I_h &: H^2(\Omega, M) \to V_h^M(G) \\
(I_h u)(x) &= u(x) \qquad \text{for all vertices } x \text{ of } G,
\end{aligned} \tag{3}$$

where $V_h^M(G)$ is the set of all geodesic finite element functions on G with values in M. When applying I_h to a function u we assume that the values of u at the vertices of G are such that the function $I_h u$ is unique. Precise conditions can be inferred from Lemma 3.2 in [5].

Interpolation errors will be estimated in terms of a generalization of the H^2 half norm suitable for functions with values in a Riemannian manifold.

Definition 3. Let $f : \Omega \to M$ be C^2. For a given coordinate system around $f(x) \in M$ denote by f^α the local coordinates and by $\Gamma^\gamma_{\alpha\beta}$ the Christoffel symbols of M. For a point $x \in \Omega$ we define (using the Einstein summation convention)

$$\alpha_x[f]^\gamma_{ij} := \partial^2_{ij} f^\gamma(x) + \Gamma^\gamma_{\alpha\beta} \partial_i f^\alpha(x) \partial_j f^\beta(x)$$

the second fundamental form of f at x.

The second fundamental form is a bilinear form on $T_x\Omega$ with values in $T_{f(x)}M$ [1].

Lemma 2. *Let g be the metric of M with components g_{ij}. The term*

$$|\alpha[f]|^2_\Omega := \int_\Omega |\alpha_x[f]|^2 \, dx \qquad \text{with } |\alpha_x[f]|^2 = g_{kl} \alpha_x[f]^k_{ij} \alpha_x[f]^l_{ij}$$

is invariant under coordinate transformations of M. It is nonnegative, and zero only if f is totally geodesic.

Proof. Invariance can be seen by direct computation. A function f is totally geodesic if and only if $\alpha_x[f] = 0$ for all $x \in \Omega$ [1]. Hence the second assertion follows.

The definition of $|\alpha[\cdot]|_\Omega$ is extended to functions in H^2 by considering the partial derivatives in a weak sense.

We can now state our main result. Remember that a triangle grid G is called quasi-uniform if there is a number $\eta > 0$ such that every triangle T of G contains a circle of radius ρ_T with $\rho_T \geq h_T/(2\eta)$, where h_T is the diameter of T [3].

Theorem 3. *Let G be a quasi-uniform simplex grid of Ω, $d \in \{1, 2\}$, and let I_h be the interpolation by geodesic finite elements defined in (3). Assume that for each element $T \in G$, the image $u(T) \subset M$ is contained in an isometric coordinate patch. Then, for each $u \in H^2(\Omega, M)$, $0 \leq m \leq 2$ we have*

$$\|\mathrm{dist}(u, I_h u)\|_m \leq C h^{2-m} |\alpha[u]|_\Omega,$$

with a constant C depending only on Ω and η.

The proof is a modification of the proof for standard finite elements given in [3]. Its main ingredient is the following local approximation result.

Lemma 3. *Let T_{ref} be the one- or two-dimensional reference simplex. For M a Riemannian manifold of zero curvature let $u : T_{ref} \to M$ be in $H^2(T_{ref}, M)$, and such that $u(T_{ref})$ is contained in a set U for which an isometric coordinate map ϕ exists. Let u_h be the geodesic interpolation function that coincides with u at the corners of T_{ref}. Then there is a positive number c such that*

$$\|\mathrm{dist}(u, u_h)\|_2 \le c|\alpha[u]|_{T_{ref}}.$$

Proof. The coordinate function ϕ is an isometry, and hence

$$\|\mathrm{dist}(u, u_h)\|_2 = \|\phi \circ u - \phi \circ u_h\|_2.$$

Since u_h is a geodesic interpolation function we can use Lemma 1 and obtain that $(\phi \circ u_h)(w) = \sum_{i=1}^{d+1} w_i \phi(u_i)$, the linear interpolation in coordinates between the values $\phi(u_i)$ at the corners of T_{ref}. Hence we can use Hilfssatz 6.2 from [3] and get

$$\|\mathrm{dist}(u, u_h)\|_2 \le c|\phi \circ u|_2 = c\sqrt{\int_{T_{\mathrm{ref}}} |D^2(\phi \circ u)|^2 \, dx},$$

where D^2 is the matrix with entries ∂^2_{ij}. Since, in the chart ϕ, the metric is the identity (Theorem 2), the Christoffel symbols vanish, and this is the coordinate expression for $|\alpha[u]|_{T_{\mathrm{ref}}}$. This proves the assertion.

Now we can prove the main approximation result.

Proof (of Theorem 3). It is sufficient to show for each triangle T_j of G the inequality

$$\|\mathrm{dist}(u, I_h u)\|_{m,T_j} \le C h^{2-m} |\alpha[u]|_{T_j} \qquad \text{for } u \in H^2(T_j, M).$$

Write $T = T_j$ for simplicity and let $\mathscr{F} : T_{\mathrm{ref}} \to T$, $\mathscr{F}(\xi) = B\xi + d$ be affine. Note that from Lemma 3 follows in particular that $|\mathrm{dist}(u, u_h)|_{l,T_{\mathrm{ref}}} \le c|\alpha[u]|_{T_{\mathrm{ref}}}$ for all $0 \le l \le 2$, where $|\cdot|_l$ is the l-th order Sobolev half norm. Use this and the integral transformation formula ([3, Formula 6.6]) to get

$$\begin{aligned}
|\mathrm{dist}(u, I_h u)|_{l,T} &\le C\|B^{-1}\|^l \cdot |\det B|^{1/2} \big|\mathrm{dist}(\mathscr{F}^{-1} \circ u, I_h(\mathscr{F}^{-1} \circ u))\big|_{m,T_{\mathrm{ref}}} \\
&\le C\|B^{-1}\|^l \cdot |\det B|^{1/2} \big|\alpha[\mathscr{F}^{-1} \circ u]\big|_{T_{\mathrm{ref}}} \\
&\le C\|B^{-1}\|^l \cdot |\det B|^{1/2} \cdot \|B\|^2 \cdot |\det B|^{-1/2} |\alpha[u]|_T \\
&\le C(\|B\| \cdot \|B^{-1}\|)^l \|B\|^{2-l} \cdot |\alpha[u]|_T.
\end{aligned}$$

Because of quasi-uniformity we have $\|B\| \cdot \|B^{-1}\| \le (2 + \sqrt{2})\eta$ and $\|B\| \le 4h$ (cf. [3]). Together we obtain

$$|\mathrm{dist}(u, I_h u)|_{l,T} \leq C h^{2-l} |\alpha[u]|_T .$$

Squaring both sides and taking the sum over l from 0 to m yields the assertion.

4 Nonlinear Scaling and the Richards Equation

As an application of the theory presented above we give a reinterpretation of a Kirchhoff-transformation-based discretization for the Richards equation in terms of geodesic finite elements. This is a surprising result, as the Richards equation and the Kirchhoff transformation are not usually associated with differential geometry. Similar results can be shown for nonlinear scaling techniques such as the one proposed by Weiser [6].

Let Ω be a domain in $\mathbb{R}^d$. The Richards equation models the evolution of a scalar pressure p in a saturated–unsaturated flow in a porous medium[2]

$$\frac{\partial}{\partial t}\theta(p) - \mathrm{div}\left(\mathrm{kr}(\theta(p))\nabla p\right) = 0.$$

The two equations of state $\theta, \mathrm{kr} : \mathbb{R} \to \mathbb{R}$ are both continuous and monotonically increasing. Implicit time discretization leads to spatial problems

$$\theta(p^n) - \tau\, \mathrm{div}\left(\mathrm{kr}(\theta(p^n))\nabla p^n\right) = \theta(p^{n-1}) \qquad \text{on } \Omega, \tag{4}$$

with n the time step number and τ the time step size.

Let G be a grid of Ω and let V_h be the space of scalar, conforming, first-order finite elements for G. Berninger et al. [2] proposed the following discretization of (4). Inserting the Kirchhoff transformation

$$p \mapsto u = \kappa(p) := \int_0^p \mathrm{kr}(\theta(q))\, dq$$

turns (4) into a semilinear problem

$$\theta(\kappa^{-1}(u^n)) - \tau \Delta u^n = \theta(\kappa^{-1}(u^{n-1})) \tag{5}$$

for a "generalized pressure" u^n. Equation (5) is equivalent to a convex minimization problem [2, Theorem 3.3]. Berninger et al. discretized it using first-order finite elements and solved the algebraic system with a monotone multigrid solver. For a discrete solution u_h of (5), a discrete physical pressure p_h was then recovered by applying the inverse discrete Kirchhoff transformation

[2] An additional term modelling the effect of gravity has been omitted for simplicity. This does not change the argument.

$$\tilde{p}_h = I_h \circ \kappa^{-1} \circ u_h \in V_h, \tag{6}$$

where I_h is the projection onto V_h by pointwise interpolation. Numerical tests showed optimal convergence orders both in the physical and the generalized pressure [2].

Note that the function $\tilde{p}_h$ from (6) is not simply the finite element solution of (4). Berninger et al. showed, however, that $\tilde{p}_h$ could be interpreted as a solution of (4) if (4) was discretized with a solution-dependent quadrature rule [2, Sect. 4.2].

We now propose a different interpretation of the solution of a Kirchhoff-transformed problem (5). Instead of using the inverse discrete Kirchhoff transform, we recover a physical pressure function with the inverse Kirchhoff transform

$$p_h = \kappa^{-1} \circ u_h \notin V_h,$$

omitting the subsequent interpolation I_h. Due to the nonlinear nature of κ, the set $V_{\kappa,h} := \kappa^{-1}(V_h)$ of functions obtained by inverse Kirchhoff transformation from first-order finite element functions is not a regular finite element space, because it does not consist of piecewise linear functions. In fact, under the usual pointwise rules for addition and scalar multiplication it does not even form a vector space.

However, $V_{\kappa,h}$ can be interpreted as a geodesic finite element space. Consider $\mathbb{R}$ as a manifold and equip it with the Riemannian metric

$$g_x(v,w) = v^T(\kappa'(x))^2 w, \qquad x \in \mathbb{R}, \quad v, w \in T_x\mathbb{R} \approx \mathbb{R},$$

which is well-defined, because κ is differentiable. Since $\mathbb{R}$ is one-dimensional it follows that $(\mathbb{R}, g)$ has zero curvature. A more instructive way to see this uses Theorem 2: The function κ is a diffeomorphism from $\mathbb{R}$ to (u_c, ∞), where $u_c = \lim_{p\to-\infty} \kappa(p) > -\infty$ is the so-called critical pressure. Hence, κ defines coordinates on the manifold $\mathbb{R}$, and we can interpret the generalized pressure u as a special coordinate on the manifold of physical pressures $\mathbb{R}$. In these coordinates the metric g is the identity

$$g_x(v,w) = ((\kappa')^{-1}v)^T(\kappa'(x))^2((\kappa')^{-1}w) = vw, \qquad x \in \mathbb{R}, \quad v, w \in T_x\mathbb{R} \approx \mathbb{R}$$

and $(\mathbb{R}, g)$ has curvature zero.

Since $(\mathbb{R}, g)$ has curvature zero we can invoke Lemma 1 to see that geodesic simplicial interpolation in the manifold $(\mathbb{R}, (\kappa')^2)$ between $d+1$ values $p_1, \ldots, p_{d+1}$ is given by

$$p_h(w) = \kappa^{-1}\Big(\sum_{i=1}^{d+1} w_i \kappa(p_i)\Big) = \kappa^{-1}\Big(\sum_{i=1}^{d+1} w_i u_i\Big) = \kappa^{-1}(u_h(w)),$$

for coordinates w on the standard simplex. This is precisely the construction of functions in $V_{\kappa,h}$ from [2] described above. We have shown the following result.

Theorem 4. *The space $V_{\kappa,h}$ is the geodesic finite element space for the manifold $\mathbb{R}$ with metric $g = (\kappa')^2$.*

This results provides a new view point on nonlinear scaling techniques.

References

1. Thierry Aubin. *Some Nonlinear Problems in Riemannian Geometry*. Springer Verlag, 1998.
2. Heiko Berninger, Ralf Kornhuber, and Oliver Sander. Fast and robust numerical solution of the Richards equation in homogeneous soil. *SIAM J. on Numerical Analysis*, 49(6):2576–2597, 2011.
3. Dietrich Braess. *Finite Elemente*. Springer Verlag, 2nd edition, 1991.
4. Halldor I. Eliasson. Geometry of manifolds of maps. *J. Differential Geometry*, 1:169–194, 1967.
5. Oliver Sander. Geodesic finite elements on simplicial grids. *Int. J. Num. Meth. Eng.*, accepted.
6. Martin Weiser. Pointwise nonlinear scaling for reaction–diffusion equations. *Appl. Num. Math.*, 59(8):1858–1869, 2009.
7. Joseph A. Wolf. *Spaces of Constant Curvature*. Publish or Perish, Inc., 3rd edition, 1974.

Design and Verification of the MPFA Scheme for Three-Dimensional Phase Field Model of Dendritic Crystal Growth

P. Strachota and M. Beneš

Abstract As an alternative to the sharp interface formulation, the phase field approach is a widely used technique for modeling phase transitions. The governing system of reaction-diffusion equations captures the instability of the underlying physical problem and is capable of modeling the evolution of complicated crystal shapes during solidification of an undercooled melt. For its numerical solution, we propose our novel anti-diffusive multipoint flux approximation (MPFA) finite volume scheme on a Cartesian mesh. The scheme is verified against the analytical solution of the modified sharp interface model. Experimental order of convergence (EOC) is measured for the temperature field in the usual norms. In addition, EOC is also obtained for the phase interface through approximating the volume of the symmetric difference of the solid phase subdomains. In the anisotropic cases including unusual higher order symmetries, computational studies with various settings also confirm convergence of our MPFA scheme which is faster than in the case of the reference finite volume scheme with 2nd order flux approximation.

1 Introduction and Background

The phase field formulation of the Stefan problem with surface tension [17] describing phase interface evolution during material solidification involves the Allen–Cahn equation [1]. In [26], the numerical solution of a problem based on this equation was used successfully in our algorithm for medical visualization. As part of this algorithm, a *multipoint flux approximation* (MPFA) finite volume scheme has been developed, proving its favorable anti-dissipative properties in visualization results. The numerical algorithm is parallel and very well scalable.

P. Strachota (✉) · M. Beneš
Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University in Prague, Prague, Czech Republic
e-mail: pavel.strachota@fjfi.cvut.cz; michal.benes@fjfi.cvut.cz

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_49,

We have therefore adapted it to the original phase field model of pure substance solidification [6] to be able to perform high resolution 3D crystal growth simulations by means of parallel computing architectures. Our very first results together with a brief description of the underlying problem have been published in [25]. This contribution provides a little more detail on the model settings (implementation of anisotropy), covers the numerical scheme design and focuses on verification of the numerical algorithm by means of measuring the convergence to a benchmark solution.

2 Problem Formulation

We start with the formulation found in [6], which describes one of the variants of the phase field model of the simplified Stefan problem. For its extension to the anisotropic case, we follow the ideas in [4]. Given a block shaped domain $\Omega \subset \mathbb{R}^3$ and time interval $\mathcal{J} = (0, T)$, the full system of phase field equations reads

$$\frac{\partial u}{\partial t} = \Delta u + L\frac{\partial p}{\partial t} \qquad \text{in } \mathcal{J} \times \Omega, \tag{1}$$

$$\alpha\xi^2\frac{\partial p}{\partial t} = \xi^2 \nabla \cdot T^0\left(\nabla p\right) + f\left(u, p, \nabla p; \xi\right) \qquad \text{in } \mathcal{J} \times \Omega, \tag{2}$$

$$u|_{t=0} = u_{ini}, \quad p|_{t=0} = p_{ini} \qquad \text{in } \Omega, \tag{3}$$

with either Dirichlet or Neumann boundary conditions. u represents the temperature field and p the phase field implicitly determining the phase interface Γ by the relation $\Gamma(t) = \left\{\mathbf{x} \in \mathbb{R}^3 \middle| p(t, \mathbf{x}) = \frac{1}{2}\right\}$. The reaction term $f\left(u, p, \nabla p; \xi\right)$ is given by

$$f\left(u, p, \nabla p; \xi\right) = ap(1-p)\left(p - \tfrac{1}{2}\right) - b\beta\xi^2 \left|\nabla p\right| \left(u - u^*\right).$$

The model parameters involve the melting point of the material u^*, the latent heat L, the attachment kinetics coefficient α, positive constants a, b, β [6] and the parameter ξ controlling the recovery of the sharp interface model [7]. The anisotropic operator T^0 (see [3, 4, 8]) is derived from the *dual* Finsler metric $\phi^0\left(\eta^*\right)$, $\eta^* \in \mathbb{R}^3$ as $T^0\left(\eta^*\right) = \phi^0\left(\eta^*\right)\phi^0_\eta\left(\eta^*\right)$ where $\phi^0_\eta = \left(\partial_{\eta_1^*}\phi^0, \partial_{\eta_2^*}\phi^0, \partial_{\eta_3^*}\phi^0\right)^{\mathrm{T}}$.

2.1 Specifying Anisotropy

The approach described in [3, 4, 7, 8] provides a general framework for treating a broad range of anisotropy settings. Putting $\phi^0\left(\eta^*\right) = \left|\eta^*\right| \psi\left(-\frac{\eta^*}{\left|\eta^*\right|}\right)$, ψ has the meaning of the anisotropic surface energy [12, 20] and assumes different forms

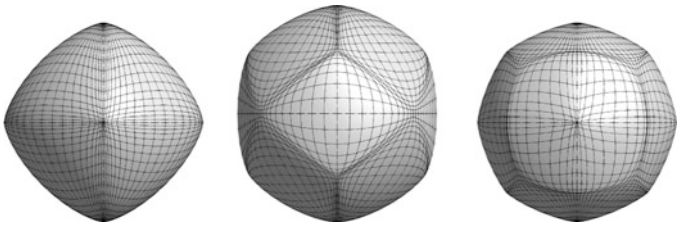


Fig. 1 Sample Wulff shapes at the limit of convexity for (from *left* to *right*) four-fold ($A_1 = 0.064$), six-fold ($A_1 = A_2 = 0.04$), and eight-fold ($A_1 = 0.0155$) anisotropy. The values of the coefficients were determined experimentally. Convex Wulff shapes with negative A_1, A_2 can also be constructed

depending on the crystalline symmetry. For the simulations, it is possible to adopt the results of [19] that provide the formulas for ψ as follows:

- Four-fold anisotropy:

$$\psi(\mathbf{n}) = 1 + A_1\left[n_1^4 + n_2^4 + n_3^4 - 6\left(n_1^2 n_2^2 + n_2^2 n_3^2 + n_3^2 n_1^2\right)\right]$$

- Six-fold anisotropy:

$$\psi(\mathbf{n}) = 1 + A_1\left(n_1^4 + n_2^4 + n_3^4 - \tfrac{3}{5}\right) + A_2\left[3\left(n_1^4 + n_2^4 + n_3^4\right) + 66 n_1^2 n_2^2 n_3^2 - \tfrac{17}{7}\right]$$

- Eight-fold anisotropy:

$$\psi(\mathbf{n}) = 1 + A_1\left[n_1^8 + n_2^8 + n_3^8 - 28\left(n_1^6 n_2^2 + n_1^2 n_2^6 + n_2^6 n_3^2 + n_2^2 n_3^6 + n_3^6 n_1^2 + n_3^2 n_1^6\right)\right] + 70 A_1\left(n_1^4 n_2^4 + n_2^4 n_3^4 + n_3^4 n_1^4\right),$$

where $\mathbf{n} = -\frac{\eta^*}{|\eta^*|}$ and the coefficients A_1, A_2 specify the anisotropy strength constrained by the requirement of strict convexity of ϕ^0 [3].

The $\phi-$unit ball $\mathscr{W}_\phi = \left\{\eta \in \mathbb{R}^3 \middle| \phi(\eta) \le 1\right\}$ represents the *Wulff shape* [12] – the equilibrium shape of the crystal. The Wulff shapes with A_1, A_2 set at the limit of convexity for the anisotropies defined above are shown in Fig. 1.

3 Numerical Solution

As in [26], the *method of lines* [22] is utilized for numerical solution. Applying a finite volume discretization scheme in space, the problem (1)–(3) is converted to a semidiscrete scheme which can be written $\forall K \in \mathscr{T}$ as

$$m(K)\frac{\mathrm{d}}{\mathrm{d}t}u_K^h(t) = \sum_{\sigma\in\mathscr{E}_K} F_{K,\sigma}(t,u_K) + L\frac{\mathrm{d}}{\mathrm{d}t}p_K^h(t), \tag{4}$$

$$\alpha\xi^2 m(K)\frac{\mathrm{d}}{\mathrm{d}t}p_K^h(t) = \xi^2\sum_{\sigma\in\mathscr{E}_K} T_{K,\sigma}^0\left(t,p_K\right) + f\left(u_K^h, p_K^h, \nabla_h p_K^h; \xi\right) \tag{5}$$

where $\mathscr{T}$ is an admissible finite volume mesh [10], $K\in\mathscr{T}$ is one particular control volume (cell) with measure $m(K)$ and $\mathscr{E}_K$ is the set of all faces of the cell K. $F_{K,\sigma}(t,u_K)$ and $T_{K,\sigma}^0\left(t,p_K\right)$ represent the respective numerical fluxes at the time t, which contain difference quotients approximating the derivatives $\partial_x w$, $\partial_y w$, $\partial_z w$ ($w\in\{u,p\}$) at the center of the face σ. To solve (4)–(5), we employ the explicit 4th order Runge–Kutta–Merson solver [9] with automatic adjustment of time step τ in agreement with the stability condition $\tau < ch^2$, $c>0$, $h=\max_{K\in\mathscr{T}}\operatorname{diam}K$.

3.1 Flux Approximation

Let us proceed to describing the particular discretizations of $\partial_x w$, $\partial_y w$, $\partial_z w$ ($w\in\{u,p\}$) in the numerical fluxes in our anti-diffusive fourth-order MPFA scheme [26]. First, the following notations need to be introduced. The computational domain Ω is a cuboid $\Omega=(0,L_1)\times(0,L_2)\times(0,L_3)$ covered by a uniform grid $\mathscr{T}$ of $n_1\times n_2\times n_3$ cells enumerated by three indices i,j,k and defined as

$$K_{i,j,k} = \left(ih_1,(i+1)h_1\right)\times\left(jh_2,\left(j+1\right)h_2\right)\times\left(kh_3,(k+1)h_3\right)$$

where $n_1,n_2,n_3\in\mathbb{N}$, $i\in\{0,1,\ldots,n_1\}$, $j\in\{0,1,\ldots,n_2\}$, $k\in\{0,1,\ldots,n_3\}$, and

$$h_1=\tfrac{L_1}{n_1},\ h_2=\tfrac{L_2}{n_2},\ h_3=\tfrac{L_3}{n_3}.$$

The grid nodes $\mathbf{x}_{K_{i,j,k}}$ are located in the center of each cell (cell-centered) according to the formula

$$\mathbf{x}_{K_{i,j,k}} = \left(\left(i+\tfrac{1}{2}\right)h_1,\left(j+\tfrac{1}{2}\right)h_2,\left(k+\tfrac{1}{2}\right)h_3\right)^{\mathrm{T}}. \tag{6}$$

For the sake of simplicity, we denote $\mathbf{x}_{i,j,k}:=\mathbf{x}_{K_{i,j,k}}$ and similarly $w_{i,j,k}^h := w_{K_{i,j,k}}^h$. We also use $w_{i,j,k}^h$ with non-integer indices to refer to interpolated values at locations specified by (6).

There are two types of difference quotients depending on the direction of the derivative with respect to the cell face:

1. For the derivative in the direction of the normal to the cell face $\sigma\in\mathscr{E}_{K_{i,j,k}}$, the node $\mathbf{x}_{i,j,k}$ and its neighbors can be used directly to assemble the stencil of the difference quotient. This is the only case appearing in $F_{K,\sigma}$.

2. For the derivative in the direction within the plane of the face $\sigma \in \mathscr{E}_{K_{i,j,k}}$, the nodes required by the difference quotient are missing and need to be interpolated from the neighboring nodes.

Consider the cell $K_{i,j,k}$ and the face $\sigma \in \mathscr{E}_{K_{i,j,k}}$ such that its outer normal vector is $\mathbf{n} = (1,0,0)^{\mathrm{T}}$ (along the x axis). The center of σ can then be denoted $\mathbf{x}_{i+\frac{1}{2},j,k}$ in agreement with (6). For the demonstration, we choose one of the two mutually perpendicular directions within the plane of σ as $\tau = (0,1,0)^{\mathrm{T}}$ (along the y axis). Then the derivatives of w are replaced by the respective difference quotients as follows:

$$\frac{\partial w}{\partial x}\left(\mathbf{x}_{i+\frac{1}{2},j,k}\right) \approx \frac{w^h_{i-1,j,k} - 27w^h_{i,j,k} + 27w^h_{i+1,j,k} - w^h_{i+2,j,k}}{24h_1}, \tag{7}$$

$$\frac{\partial w}{\partial y}\left(\mathbf{x}_{i+\frac{1}{2},j,k}\right) \approx \frac{w^h_{i+\frac{1}{2},j-2,k} - 8w^h_{i+\frac{1}{2},j-1,k} + 8w^h_{i+\frac{1}{2},j+1,k} - w^h_{i+\frac{1}{2},j+2,k}}{12h_2} \tag{8}$$

where for $\tilde{j} \in \{j-2, j-1, j+1, j+2\}$

$$w^h_{i+\frac{1}{2},\tilde{j},k} = \tfrac{1}{16}\left(-w^h_{i-1,\tilde{j},k} + 9w^h_{i,\tilde{j},k} + 9w^h_{i+1,\tilde{j},k} - w^h_{i+2,\tilde{j},k}\right). \tag{9}$$

Note that in (7), the used stencil is not equidistant. The cubic interpolation (9) guarantees fourth order of approximation in (8) provided that the function w is smooth enough.

3.2 *Discretization of Other Terms*

The remaining differential terms in (4)–(5) are discretized as follows:

1. The gradient ∇p in the reaction term of (2) is approximated by a "discrete gradient" $\nabla_h p^h_K$ in (4). Its components are calculated by a 4-th order equidistant stencil at the cell centers. For $K = K_{i,j,k}$, the first component reads

$$\frac{\partial p}{\partial x}\left(\mathbf{x}_{i,j,k}\right) \approx \frac{p^h_{i-2,j,k} - 8p^h_{i-1,j,k} + 8p^h_{i+1,j,k} - p^h_{i+1,j,k}}{12h_1}. \tag{10}$$

2. The term $\frac{\mathrm{d}}{\mathrm{d}t} p^h_K(t)$ in (4) is expressed from (5) in a straightforward manner.

4 Convergence Studies

To verify the numerical scheme, *experimental order of convergence* (EOC) [21] computations have been performed. The numerical solution has been found on a sequence of gradually refining grids, making it possible to calculate the ratio

Table 1 EOC results for the temperature field. $\Omega = (0,4)^3$, $T = 0.75$, sampling in 20 time levels

h	$L_\infty(\mathscr{J}; L_2(\Omega))$ error $\times 10^{-2}$	EOC in $L_\infty(\mathscr{J}; L_2(\Omega))$	$L_\infty(\mathscr{J}; L_\infty(\Omega))$ error $\times 10^{-2}$	EOC in $L_\infty(\mathscr{J}; L_\infty(\Omega))$
4/100	32.6991	–	8.5080	–
4/200	9.3847	1.8009	3.9355	1.1123
4/300	5.6551	1.2492	2.2397	1.3902
4/400	4.0763	1.1379	1.7180	0.9218

Table 2 EOC results for the phase interface. $\Omega = (0,4)^3$, $T = 0.75$, sampling in 20 time levels

h	$L_\infty(\mathscr{J}; SSDV)$ error $\times 10^{-1}$	EOC in $L_\infty(\mathscr{J}; SSDV)$
4/100	6.5798	–
4/200	5.6649	0.2160
4/300	3.8409	0.9584
4/400	2.8403	1.0491

$$\mathrm{EOC}_i = \log\left(\frac{\mathrm{Error}_i}{\mathrm{Error}_{i-1}}\right) / \log\left(\frac{h_i}{h_{i-1}}\right)$$

where $h = \max_K \mathrm{diam}(K)$ is the mesh size and Error_i is the difference of the i-th solution from the analytical benchmark solution measured in an appropriate norm. The analytical solution has been adopted from [23] where, however, an isotropic *sharp interface* model is discussed. Therein are provided the formulas for the evolution of the temperature field u and the phase interface Γ.

To observe convergence to the sharp interface solution, the parameter ξ of the phase field model needs to tend to zero together with grid refinement. As ξ determines the thickness of the diffuse interface [5], the condition $h < C\xi$, $C > 0$ applies to ensure sufficient mesh resolution at the interface. Our simulations indicate that $C \in (1,4)$ depending on model parameters. For the EOC measurement, we have fixed the ratio $h = \frac{10}{11}\xi$. It is then easy to evaluate EOC for the temperature field in any of the suitable norms, as in [23] and also e.g. in [8]. On top of that, we have also successfully measured the convergence of the interface $\Gamma^h(t) = \left\{\mathbf{x} \in \Omega \,\middle|\, p^h(t,\mathbf{x}) = \frac{1}{2}\right\}$ found by the numerical simulation to the sharp interface Γ by estimating the *symmetric set difference volume*

$$SSDV(A,B) = m\left[\left(A \backslash B\right) \cup \left(B \backslash A\right)\right] = m\left(A \backslash B\right) + m\left(B \backslash A\right)$$

where $A = \mathrm{int}\,\Gamma^h(t) = \left\{\mathbf{x} \in \Omega \,\middle|\, p^h(t,\mathbf{x}) > \frac{1}{2}\right\}$, $B = \mathrm{int}\,\Gamma(t)$. The results are summarized in Table 1 for the temperature field and in Table 2 for the phase interface. They both confirm 1st order convergence of the phase field model to the sharp interface formulation and at least 1st order convergence of the numerical scheme to the precise solution of the phase field model (2)–(3).

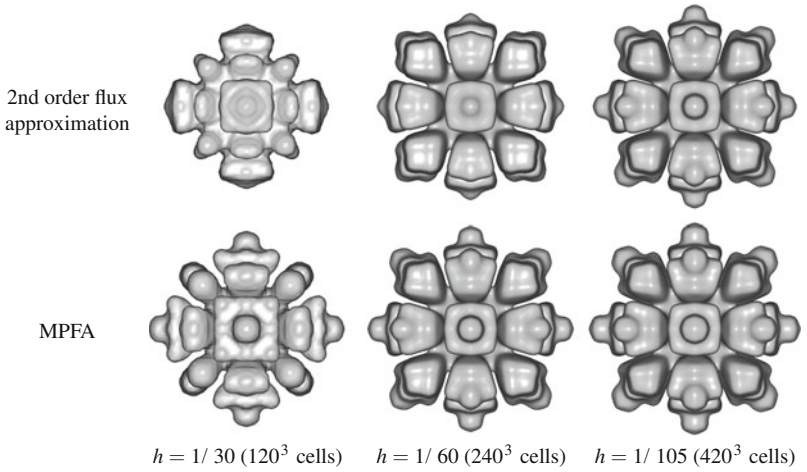


Fig. 2 Visual convergence study of different flux approximation schemes for the case of eight-fold anisotropy and fixed $\xi = 1.05/105$. With decreasing h, results of both schemes approach the same limit shape, but the MPFA scheme converges faster. Other model parameters: $A_1 = 0.01$, $\Omega = (0, 4)^3$, $a = 2$, $b = 1$, $\alpha = 3$, $\beta = 300$, $L = 2$, $u^* = 1$, $T = 0.1$

5 Convergence in Anisotropic Cases

Up to the authors' knowledge, there is no analytical solution available for comparison in the anisotropic settings. Also, we are interested in the performance of the numerical solver in the cases of complex patterns formation, which no analytical solution can describe. Our conclusions are therefore based on visual observation only. To assess the performance of the numerical scheme, we compare its results to those obtained by the well known 2nd order central difference flux approximation scheme. Most tests such as that in Fig. 2 imply that the crystal shape converges with grid refinement and that the rate of convergence is faster for the MPFA scheme.

6 Conclusion

Understanding material solidification and crystal growth is important in many engineering and industrial applications. A lot of existing results of numerical simulations of these processes are limited to 2D [15,16,18] or rely on simplifications such as shape symmetries to compensate for the lack of computational resources [11, 13, 14, 23, 24]. Moreover, even though some exceptions appeared recently [2], the 3D simulations usually only consider four-fold anisotropy. Contrary to that, we have developed a parallel numerical algorithm capable of high resolution 3D crystal growth modeling with a variety of anisotropies naturally incorporated

into the model. The convergence of the used numerical schemes has been verified experimentally. In the simulations with anisotropy, the anti-diffusive MPFA scheme seems to treat evolution of complex dendritic structures more accurately.

Acknowledgements This work was partially supported by the following projects: The project of the Ministry of Education of the Czech Republic MSM6840770010 *Applied Mathematics in Technical and Physical Sciences*. The Grant Agency of the Czech Technical University in Prague, grant No. SGS11/161/OHK4/3T/14.

References

1. Allen, S., Cahn, J.W.: A microscopic theory for antiphase boundary motion and its application to antiphase domain coarsening. Acta Metall. **27**, 1084–1095 (1979)
2. Barrett, J.W., Garcke, H., Nürnberg, R.: On stable parametric finite element methods for the Stefan problem and the Mullins-Sekerka problem with applications to dendritic growth. J. Comput. Phys. **229**, 6270–6299 (2010)
3. Bellettini, G., Paolini, M.: Anisotropic motion by mean curvature in the context of Finsler geometry. Hokkaido Math. J. **25**(3), 537–566 (1996)
4. Beneš, M.: Anisotropic phase-field model with focused latent-heat release. In: FREE BOUNDARY PROBLEMS: Theory and Applications II, *GAKUTO International Series in Mathematical Sciences and Applications*, vol. 14, pp. 18–30 (2000)
5. Beneš, M.: Mathematical analysis of phase-field equations with numerically efficient coupling terms. Interface. Free. Bound. **3**, 201–221 (2001)
6. Beneš, M.: Mathematical and computational aspects of solidification of pure substances. Acta Math. Univ. Comenianae **70**(1), 123–151 (2001)
7. Beneš, M.: Diffuse-interface treatment of the anisotropic mean-curvature flow. Appl. Math-Czech. **48**(6), 437–453 (2003)
8. Beneš, M.: Computational studies of anisotropic diffuse interface model of microstructure formation in solidification. Acta Math. Univ. Comenianae **76**, 39–59 (2007)
9. Butcher, J.C.: Numerical Methods for Ordinary Differential Equations. Wiley (2003)
10. Eymard, R., Gallouët, T., Herbin, R.: Finite volume methods. In: P.G. Ciarlet, J.L. Lions (eds.) Handbook of Numerical Analysis, vol. 7, pp. 715–1022. Elsevier (2000)
11. Green, J.R., Jimack, P.K., Mullis, A.M., Rosam, J.: An adaptive, multilevel scheme for the implicit solution of three-dimensional phase-field equations. Numer. Meth. Part. D. E. **27**, 106–120 (2010)
12. Gurtin, M.E.: Thermomechanics of Evolving Phase Boundaries in the Plane. Oxford Mathematical Monographs. Oxford University Press (1993)
13. Karma, A., Rappel, W.J.: Numerical simulation of three-dimensional dendritic growth. Phys. Rev. Lett. **77**(19), 4050–4053 (1996)
14. Karma, A., Rappel, W.J.: Quantitative phase-field modeling of dendritic growth in two and three dimensions. Phys. Rev. E **57**(4), 4 (1998)
15. Kupferman, R., Shochet, O., Ben-Jacob, E.: Numerical study of a morphology diagram in the large undercooling limit using a phase-field model. Phys. Rev. E. **50**(2), 1005–1008 (1993)
16. McFadden, G.B., Wheeler, A.A., Braun, R.J., Coriell, S.R.: Phase-field models for anisotropic interfaces. Phys. Rev. E **48**(3), 2016–2024 (1993)
17. Meirmanov, A.M.: The Stefan Problem. De Gruyter Expositions in Mathematics. Walter de Gruyter (1992)
18. Mullis, A.M., Cochrane, R.F.: A phase field model for spontaneous grain refinement in deeply undercooled metallic melts. Acta Mater. **49**, 2205–2214 (2001)

19. PunKay, M.: Modeling of anisotropic surface energies for quantum dot formation and morphological evolution. In: NNIN REU Research Accomplishments, pp. 116–117. University of Michigan (2005)
20. R. E. Napolitano, S.L.: Three-dimensional crystal-melt Wulff-shape and interfacial stiffness in the Al-Sn binary system. Phys. Rev. B **70**(21), 214,103 (2004)
21. Rice, J.R., Mu, M.: An experimental performance analysis for the rate of convergence of 5-point star on general domains. Tech. rep., Department of Computer Sciences, Purdue University (1988)
22. Schiesser, W.E.: The Numerical Method of Lines: Integration of Partial Differential Equations. Academic Press, San Diego (1991)
23. Schmidt, A.: Computation of three dimensional dendrites with finite elements. J. Comput. Phys. **125**, 293–3112 (1996)
24. Singer, H.M., Singer-Loginova, I., Bilgam, J.H., Amberg, G.: Morphology diagram of thermal dendritic solidification by means of phase-field models in two and three dimensions. J. Cryst. Growth. **296**, 58–68 (2006)
25. Strachota, P., Beneš, M.: A multipoint flux approximation finite volume scheme for solving anisotropic reaction-diffusion systems in 3D. In: J. Fořt, J. Fürst, J. Halama, R. Herbin, F. Hubert (eds.) Finite Volumes for Complex Applications VI - Problems & Perspectives, pp. 741–749. Springer (2011). DOI 10.1007/978-3-642-20671-9_78
26. Strachota, P., Beneš, M., Tintěra, J.: Towards clinical applicability of the diffusion-based DT-MRI visualization algorithm. J. Vis. Commun. Image R. **23**(2), 387–396 (2012). DOI 10.1016/j.jvcir.2011.11.009

An Evolving Surface Finite Element Method for the Numerical Solution of Diffusion Induced Grain Boundary Motion

V. Styles

Abstract We apply an evolving surface finite element method (ESFEM) to a mathematical model for diffusion induced grain boundary motion. The model involves the coupling of a diffusion equation on a moving surface to an equation for the motion of the surface. We formulate a finite element approximation of the model which involves triangulated surfaces whose vertices move in time. We present numerical simulations.

1 Introduction

We consider the problem of determining a hypersurface $\Gamma(t) \subset \mathbb{R}^3$ that evolves with the velocity law

$$V = \kappa + \alpha c^2 \tag{1a}$$

where c is a scalar function that solves the diffusion equation

$$\partial^\circ c \ = D\Delta_\Gamma c - c\nabla_\Gamma \cdot V\boldsymbol{\nu} - \beta V c \quad \text{on } \Gamma(t). \tag{1b}$$

Here $V > 0$ and κ respectively denote the normal velocity and the mean curvature of $\Gamma(t)$, $\partial^\circ c = \frac{\partial c}{\partial t} + V\boldsymbol{\nu} \cdot \nabla c$ denotes the material derivative of c along flow lines orthogonal to $\Gamma(t)$, $D > 0$ is a constant diffusivity parameter and $\Delta_\Gamma c = \nabla_\Gamma \cdot \nabla_\Gamma c$ is the Laplace Beltrami operator (or surface Laplacian), with $\nabla_\Gamma c = \nabla c - \nabla c \cdot \boldsymbol{\nu}\boldsymbol{\nu}$ denoting the tangential gradient and $\boldsymbol{\nu}$ the unit normal to $\Gamma(t)$. Lastly α and β are positive physical constants. This model can be used to describe the physical

V. Styles (✉)
University of Sussex, Brighton BN1 9QH, UK
e-mail: v.styles@sussex.ac.uk

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_50,

phenomenon of diffusion induced grain boundary motion, which is the motion of grain boundaries in thin metallic films due to the absorption of solute from an external vapour, [4, 13]. Here the surface $\Gamma(t)$ represents the grain boundary and $c(x,t)$ is a scalar function that denotes the concentration of solute on $\Gamma(t)$.

There are three main techniques for solving geometric evolution equations of the form (1a); the parametric approach, the level set method and the phase field approach, see [6]. Related to these three techniques are recently introduced models for solving advection-diffusion partial differential equations on evolving surfaces of the form (1b); the evolving surface finite element method proposed in [8], an Eulerian surface finite element method, [9], and a diffuse interface model [10, 11, 15]. Here we consider the parametric approach coupled with the evolving surface finite element method. In particular we use techniques introduced in [2] to present a parametric finite element approximation of the hypersurface $\Gamma \in \mathbb{R}^3$ moving with the forced geometric motion law (1a). This approximation gives rise to triangulated surfaces Γ_h^n on which (1b) needs to be approximated. To this end we use the evolving surface finite element method derived in [8] whereby a finite element space is defined that is the space of continuous piecewise linear functions on the triangulated surface.

The free boundary problem (1a) and (1b) arises from formal asymptotics on the phase field model for diffusion induced grain boundary motion presented in [4] and existence and uniqueness of classical solutions to this free boundary problem are presented in [14].

The structure of the article is as follows. In the next section we introduce a weak formulation of problem (1a) and (1b) then in Sect. 3 we present a finite element approximation of the model. We conclude with Sect. 4 in which we present some numerical simulations.

2 The Model

In this section we describe the geometrical configuration that we consider and we present a weak formulation of the model.

2.1 Geometrical Configuration

The geometrical configuration we study takes the form of a domain $\Omega := (-1,1) \times (0,L) \times (-1,1)$, containing a single hypersurface $\Gamma(t)$ that spans the height (x_3 direction) and width (x_1 direction) of the domain. We assume that $\Gamma(t)$ never comes in contact with the planes $x_2 = 0$ or $x_2 = L$. We supplement (1a) and (1b) with the following boundary data

$$c(x,t) = c_+ \ \forall x \in \Gamma(t) \cap \{x_3 = 1\}, \quad c(x,t) = c_- \ \forall x \in \Gamma(t) \cap \{x_3 = -1\}, \tag{2a}$$

where c_+ and c_- are positive constants, and we impose the natural boundary condition

$$\nabla c \cdot \boldsymbol{\mu} = 0 \ \ \forall x \in \Gamma(t) \cap \{x_1 = \pm 1\}, \tag{2b}$$

where $\boldsymbol{\mu}$ is tangential to $\partial\Gamma(t)$ and normal to $\partial\Omega$. Furthermore we set

$$c(x,0) = 0 \ \ \forall x \in \Gamma(0). \tag{2c}$$

Physically the conditions (2a)–(2c) imply that solute is only absorbed into the grain boundary from the top and bottom of the film and that initially there is no solute in the film. To supplement (1a) we set $\Gamma(0) = \Gamma_0$ and we impose that the surface remains orthogonally attached to the boundaries $x_1 = \pm 1$ and $x_3 = \pm 1$ of Ω.

2.2 Weak Formulation of the Model

Here we introduce a weak formulation of the model. First we give a parametric formulation of (1a): for parametrizations $\mathbf{x} : \Sigma \times [0,T] \to \mathbb{R}^3$, with $\mathbf{x}(\cdot,0) = \mathbf{x}_0(\cdot)$, where Σ is a suitable compact reference manifold in $\mathbb{R}^3$, (1a) can be written in the form

$$V := \mathbf{x}_t \cdot \boldsymbol{\nu} = \kappa + \alpha c^2, \quad \kappa \boldsymbol{\nu} = \Delta_\Gamma \mathbf{x}, \tag{3}$$

where the second identity in (3) was used for the first time by Dziuk in [7] in designing a finite element method for mean curvature flow. Second we follow the techniques introduced by Dziuk and Elliott in [8] and introduce a weak formulation of (1b). Multiplying (1b) by $\phi \in W(\Gamma(t))$, where

$$W(\Gamma(t)) = \{\eta \in H^1(\Gamma(t)) : \eta(x,t) = 0 \ \ \forall x \in \Gamma(t) \cap \{x_3 = \pm 1\},$$

integrating over $\Gamma(t)$ and integrating by parts yields

$$\int_{\Gamma(t)} \left(\partial^\circ c \ \phi + D \nabla_\Gamma c \cdot \nabla_\Gamma \phi + c\phi \nabla_\Gamma \cdot V\boldsymbol{\nu} + \beta V c \phi \right) = 0 \quad \forall \phi \in W(\Gamma(t)). \tag{4}$$

The function c is defined to be a weak solution of (1b) if (4) holds for almost every $t \in (0,T)$.

We now note a transport formula see [8], which states that if f is a function defined in a neighbourhood of a surface $\Gamma(t)$ that is evolving with velocity $\mathbf{v} = V\boldsymbol{\nu}$, then

$$\frac{d}{dt} \int_{\Gamma(t)} f = \int_{\Gamma(t)} \partial^\circ f \ + f \ \nabla_\Gamma \cdot \mathbf{v}. \tag{5}$$

Using (5) we can reformulate (4) as

$$\frac{d}{dt}\int_{\Gamma(t)} c\phi + \int_{\Gamma(t)} \left(D\nabla_\Gamma c\cdot\nabla_\Gamma\phi + \beta V c\phi\right) = \int_{\Gamma(t)} c\partial^\circ\phi \quad \forall\phi\in W(\Gamma(t)).$$

Thus we arrive at the following weak formulation of the model

$$\mathbf{x}(\cdot,0) := \mathbf{x}_0(\cdot), \tag{6a}$$

$$\mathbf{x}_t\cdot\boldsymbol{\nu} = \kappa + \alpha c^2, \quad \kappa\boldsymbol{\nu} = \Delta_\Gamma\mathbf{x}, \tag{6b}$$

$$\text{the surface remains orthogonally attached to } \partial\Omega\cap\{x_1=\pm1\cup x_3=\pm1\}, \tag{6c}$$

$$\frac{d}{dt}\int_{\Gamma(t)} c\phi + \int_{\Gamma(t)} \left(D\nabla_\Gamma c\cdot\nabla_\Gamma\phi + \beta V c\phi\right) = \int_{\Gamma(t)} c\partial^\circ\phi \quad \forall\phi\in W(\Gamma(t)), \tag{6d}$$

$$c(x,0) = 0 \quad \forall x\in\Gamma(0), \tag{6e}$$

$$c(x,t) = c_+ \quad \forall x\in\Gamma(t)\cap\{x_3=1\},\ c(x,t)=c_- \quad \forall x\in\Gamma(t)\cap\{x_3=-1\},\ t>0. \tag{6f}$$

3 Finite Element Discretization

In this section we introduce some notation and then we present a finite element discretization of the model (6a)–(6f).

3.1 Notation

For a continuous in time discretization of (6a)–(6f) we approximate $\Gamma(t)$ by a triangulated evolving surface $\Gamma_h(t)$, such that $\Gamma_h(t) = \cup_{j=1}^J \overline{\sigma_j(t)}$ where $\{\sigma_j(t)\}_{j=1}^J$ is a family of mutually disjoint open triangles. We denote the vertices of $\sigma_j(t)$ by $\{\mathbf{q}_{j_k}(t)\}_{k=0}^2$ and we define the unit normal $\boldsymbol{\nu}(t)$ to $\Gamma_h(t)$ such that

$$\boldsymbol{\nu}_j(t) := \boldsymbol{\nu}|_{\sigma_i(t)} := \frac{(\mathbf{q}_{j_1}(t)-\mathbf{q}_{j_0}(t))\times(\mathbf{q}_{j_2}(t)-\mathbf{q}_{j_0}(t))}{|(\mathbf{q}_{j_1}(t)-\mathbf{q}_{j_0}(t))\times(\mathbf{q}_{j_2}(t)-\mathbf{q}_{j_0}(t))|} \quad \text{for } j=1\to J.$$

We define $\Gamma_h^\pm(t) := \Gamma_h(t)\cap\{x_3=\pm1\}$, and let I be the number and $\mathscr{I}$ the set of vertex indices, such that $\mathscr{I} := \mathscr{I}_B^+\cup\mathscr{I}_B^-\cup\mathscr{I}_I$, where $\mathscr{I}_B^\pm$ denotes the set of nodes that lie on $\Gamma_h^\pm(t)$. For each t we define the finite element space

$$S^h(\Gamma_h(t)) = \{\chi \in C(\Gamma_h(t))|\ \chi|_{\sigma_j} \text{ is piecewise linear for } j = 1 \to J\}$$

with $\{\chi_i\}_{i=1}^I$ denoting the standard basis of $S^h(t)$. We set

$$S_0^h(\Gamma_h(t)) := \{\chi \in S^h(\Gamma_h(t))|\ \chi = 0 \ \text{ on } \partial\Gamma_h(t) \cap \{x_3 = \pm 1\}\}$$

and

$$S_b^h(\Gamma_h(t)) := \{\chi \in S^h(\Gamma_h(t))|\ \chi = c_\pm \text{ on } \partial\Gamma_h^\pm, \text{ and } \nabla\chi\cdot\boldsymbol{\mu} = 0 \text{ on } \partial\Gamma_h(t)\cap\{x_1 = \pm 1\}\}.$$

Since the surface $\Gamma(t)$ intersects Ω we follow the techniques used in [3, 5] and define

$$\underline{Z}^h(\Gamma_h(t)) := \{\boldsymbol{\chi} \in [S^h(\Gamma_h(t))]^3|\ \chi_1 = 0 \text{ on } \partial\Gamma_h \cap \{x_1 = \pm 1\},\ \chi_3 = 0 \text{ on } \partial\Gamma_h^\pm\},$$

and

$$\underline{Z}_b^h(\Gamma_h(t)) := \{\boldsymbol{\chi} \in [S^h(\Gamma_h(t))]^3|\ \chi_1 = \pm 1 \text{ on } \partial\Gamma_h \cap \{x_1 = \pm 1\},\ \chi_3 = \pm 1 \text{ on } \partial\Gamma_h^\pm\}.$$

Next we follow the authors in [2] and introduce a weighted normal, $\boldsymbol{\omega}(t) := \sum_{i=1}^I \boldsymbol{\omega}_i(t)\chi_i$ such that $\boldsymbol{\omega}_i(t)$ can be interpreted as a weighted normal defined at the node $\mathbf{q}_i(t)$ of the surface $\Gamma_h(t)$ and is defined by

$$\boldsymbol{\omega}_i(t) := \frac{1}{|\Lambda_i(t)|} \sum_{\sigma_j(t)\in\mathscr{T}_h(t)} |\sigma_j(t)|\boldsymbol{\nu}_j(t)$$

where $\mathscr{T}_h(t) := \{\sigma_j(t) : \mathbf{q}_i(t) \in \overline{\sigma_j(t)}\}$, $\Lambda_i(t) := \cup_{\sigma_j(t)\in\mathscr{T}_h(t)}\overline{\sigma_j(t)}$ and $|\sigma_j(t)|$ is the measure of $\sigma_j(t)$.

For the fully discrete discretization we set $t_m = m\tau$, $m = 0 \to M$ and for each t_m, $m = 0 \to M$, we define $\Gamma_h^m := \Gamma_h(t_m)$, $\sigma_j^m := \sigma_j(t_m)$ and $\boldsymbol{\omega}^m := \boldsymbol{\omega}(t_m)$. Following [2] for scalar (and vector) functions $u, v \in L^2(\Gamma)$ $(\boldsymbol{u}, \boldsymbol{v} \in [L^2(\Gamma)]^3)$ we introduce the L^2 inner product $\langle\cdot,\cdot\rangle_m$ over Γ_h^m: $\langle u, v\rangle_m := \int_{\Gamma_h^m} u\cdot v$ and for piecewise continuous functions u, v we introduce the mass lumped inner product $\langle\cdot,\cdot\rangle_m^h$:

$$\langle u, v\rangle_m^h := \frac{1}{3}\sum_{j=1}^J |\sigma_j^m| \sum_{k=0}^2 (u\cdot v)((\mathbf{q}_{j_k}^m)^-)$$

where $u((\mathbf{q}_{j_k}^m)^-) := \lim_{\sigma_j^m \ni \mathbf{p} \to q_{j_k}^m} u(\mathbf{p})$.

3.2 A Fully Discrete Finite Element Approximation of (6b)

We use the approach of Barrett, Garcke and Nürnberg presented in [2], to give a finite element approximation of (6b).

Given a parametrization $\mathbf{X}^{m-1} \in \underline{Z}_b^h(\Gamma_h^{m-1})$ of Γ_h^{m-1} and an approximation $C_h^{m-1} \in S_b^h(\Gamma_h^{m-1})$ to $c(t_{m-1})$, find $\{\mathbf{X}^m, \kappa^m\} \in \underline{Z}_b^h(\Gamma_h^{m-1}) \times S^h(\Gamma_h^{m-1})$ such that

$$\frac{1}{\tau}\langle \mathbf{X}^m - \mathbf{X}^{m-1}, \chi \boldsymbol{\nu}^{m-1}\rangle_{m-1}^h - \langle \kappa^m, \chi\rangle_{m-1}^h = \alpha \langle (C_h^{m-1})^2, \chi\rangle_{m-1}^h \ \forall \chi \in S^h(\Gamma_h^{m-1}) \tag{7a}$$

$$\langle \kappa^m \boldsymbol{\nu}^m, \boldsymbol{\chi}\rangle_{m-1}^h + \langle \nabla_{\Gamma_h^{m-1}} \mathbf{X}^m, \nabla_{\Gamma_h^{m-1}} \boldsymbol{\chi}\rangle_{m-1} = 0 \quad \forall \boldsymbol{\chi} \in \underline{Z}^h(\Gamma_h^{m-1}). \tag{7b}$$

3.3 Semi-discrete Finite Element Approximation of (6d)

In order to approximate the diffusion equation (6d) we use the evolving surface finite element method introduced by Dziuk and Elliott in [8]. We begin with the following continuous in time approximation of (6d): Find $C_h(\cdot, t) \in S_b^h(\Gamma_h(t))$ such that

$$\frac{d}{dt}\int_{\Gamma_h(t)} C_h \chi + \int_{\Gamma_h(t)} \left(D \nabla_{\Gamma_h} C_h \cdot \nabla_{\Gamma_h} \chi + \beta V_h C_h \chi \right) = \int_{\Gamma_h(t)} C_h \partial^\circ \chi \quad \forall \chi \in S_0^h(\Gamma_h(t)). \tag{8}$$

Here $V_h(\cdot, t)$ denotes the normal velocity of Γ_h and is given by $V_h(\cdot, t) = \sum_{i=1}^{I} V_i(t)\chi_i(\cdot, t)$ with $V_i(t) = \frac{d\mathbf{X}_i}{dt}(t) \cdot \boldsymbol{\omega}_i(t)$.

Recalling that the nodal basis functions of $S^h(\Gamma_h(t))$ are denoted by $\{\chi_i(\cdot, t)\}_{i=1}^I$, from [8] we have that if the nodes move with a velocity $\mathscr{V} = V\boldsymbol{\nu} + \mathbf{T}$ then the basis functions satisfy the transport property

$$0 = \partial^\bullet \chi_i := \partial^\circ \chi_i + \mathbf{T} \cdot \nabla \chi_i \quad \text{for } i = 1 \to I$$

$$\Rightarrow \ \partial^\circ \chi_i = -\mathbf{T} \cdot \nabla \chi_i \quad \text{for } i = 1 \to I. \tag{9}$$

Remark 1. From (9) we note that if the velocity of the nodes is orthogonal to Γ_h then we have $\partial^\circ \chi_i = \partial^\bullet \chi_i = 0$, for $i = 1 \to I$ and (8) reduces to

$$\frac{d}{dt}\int_{\Gamma_h(t)} C_h \chi_i + \int_{\Gamma_h(t)} \left(D \nabla_{\Gamma_h} C_h \cdot \nabla_{\Gamma_h} \chi_i + \beta V_h C_h \chi_i \right) = 0 \ \ \forall\, i \in \mathscr{I}_I. \tag{10}$$

3.4 *Fully-Discrete Finite Element Approximation of (6d)*

For a fully discrete approximation of (6d) we use a semi implicit time discretization; setting C_h^m to represent $C_h(\cdot, t_m)$ and noting (8) and (9) we have:

Given Γ_h^{m-1}, Γ_h^m and $C_h^{m-1} \in S_b^h(\Gamma_h^{m-1})$, find $C_h^m \in S_b^h(\Gamma_h^m)$ such that for all $i \in \mathscr{I}_I$

$$\frac{1}{\tau}\langle C_h^m, \chi_i^m\rangle_m^h - \frac{1}{\tau}\langle C_h^{m-1}, \chi_i^{m-1}\rangle_{m-1}^h + D\langle \nabla_{\Gamma_h^m} C_h^m, \nabla_{\Gamma_h^m}\chi_i^m\rangle_m + \beta\langle V_h^m C_h^m, \chi_i^m\rangle_m^h + \langle C_h^m, \mathbf{T}_h^m \cdot \nabla_{\Gamma_h^m}\chi_i^m\rangle_m^h = 0. \quad (11a)$$

Here $V_h^m = \sum_{i=1}^I V_i^m \chi_i^m$ denotes the fully-discrete normal velocity of Γ_h^m and $\mathbf{T}_h^m = \sum_{i=1}^I \mathbf{T}_i^m \chi_i^m$ denotes the fully-discrete tangential velocity of Γ_h^m, with $V_i^m := \frac{1}{\tau}(\mathbf{X}_i^m - \mathbf{X}_i^{m-1}) \cdot \boldsymbol{\omega}_i^m$ and $\mathbf{T}_i^m := \frac{1}{\tau}\left([\mathbf{X}_i^m - \mathbf{X}_i^{m-1}] - [\mathbf{X}_i^m - \mathbf{X}_i^{m-1}] \cdot \boldsymbol{\omega}_i^m \boldsymbol{\omega}_i^m\right)$. Discretising (6e-f) gives

$$C_i^0 = 0 \quad \text{for } i \in \mathscr{I} \quad \text{and} \quad C_i^m = c_\pm \ \text{ for } i \in \mathscr{I}_B^\pm, \quad m = 1 \to M. \quad (11b)$$

4 Numerical Results

In this section we display numerical simulations obtained from the scheme (7a), (7b) and (11a), (11b). All the simulations presented were produced using the finite element toolbox ALBERTA, [16] and visualised using the visualisation application PARAVIEW, [1].

We show two sets of results. In both sets we set $\Omega = (-1, 1)\times(-0.1, 4)\times(-1, 1)$, and we set $\Gamma(0)$ to be the planar surface $x_2 = 0$ with $C^0(x) \equiv 0$. Furthermore we set $D = 1$, $\alpha = 5$, β =5,000. In the first set of results, Fig. 1, we set $c_+ = c_- = 1$, while in the second set, Fig. 2, we set $c_+ = 1$ and $c_- = 0.5$.

In Fig. 1 the four subplots display the approximate solution $C_h(t_m)$ on $\Gamma_h(t_m)$ at times $t_m = 0.2, 0.5, 0.8, 1.1$. Since the interface is close to planar during the early stages of motion the concentration term in (1a) dominates the motion and as the concentration of solute is higher at the top and bottom of the interface these parts of Γ_h move faster than the middle section. Thus the mean curvature, κ, of Γ becomes larger in the middle of the domain resulting in this part of Γ_h now moving faster than the parts at the top and bottom. The consequence of this motion is that after some time the concentration distribution and the shape of Γ_h do not change (see the final two subplots). In particular a travelling wave solution, of the kind studied in [12], has been reached.

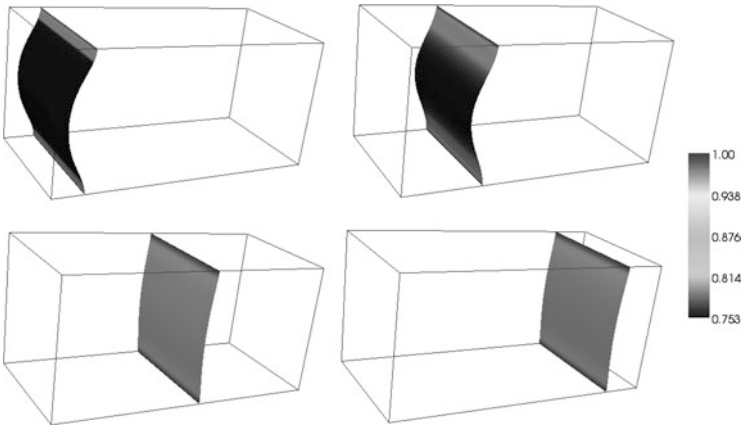


Fig. 1 Evolution of a planar surface with $c_+ = 1,\ c_- = 1,\ D = 1,\ \alpha = 5,\ \beta$ =5,000

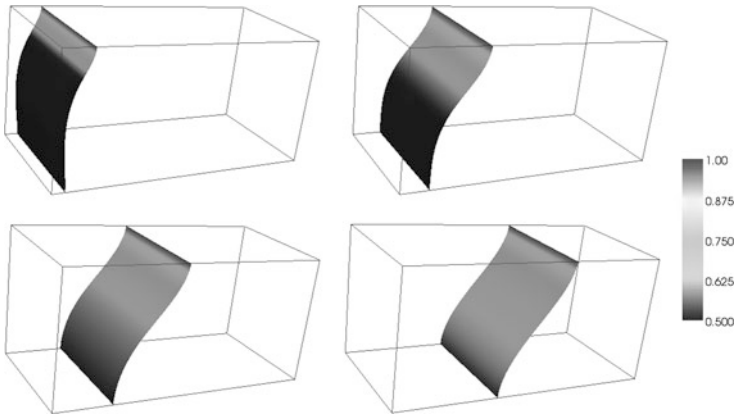


Fig. 2 Evolution of a planar surface with $c_+ = 1,\ c_- = 0.5,\ D = 1,\ \alpha = 5,\ \beta$ =5,000

In Fig. 2 we diplay the approximate solution $C_h(t_m)$ on $\Gamma_h(t_m)$ at times $t_m = 0.2, 0.5, 0.8, 1.0$. In these simulations the concentration of solute that diffuses in from the top is set to be twice that which diffuses from the bottom and as a result the top of the interface always moves faster than the bottom. Again the problem reaches a travelling wave solution.

Acknowledgements This work was supported by the EPSRC grant EP/D078334/1.

References

1. HENDERSON, A., (2007), *ParaView Guide, A Parallel Visualization Application*, Kitware Inc.
2. BARRETT, J.W., GARCKE, H. & NÜRNBERG, R., (2008), *On the parametric finite element approximation of evolving hypersurfaces in* $\mathbb{R}^3$, J. Comput. Phys., **227**, 4281–4307.
3. BARRETT, J.W., GARCKE, H. & NÜRNBERG, R., (2007), *On the variational approximation of combined second and fourth order geometric evolution equations*, SIAM J. Scientific Comp., **29**, 1064–8275.
4. CAHN, J. W., FIFE, P. & PENROSE, O., (1997), *A phase-field model for diffusion-induced grain-boundary motion*, Acta Mater., **45**, 4397–4413.
5. DECKELNICK, K. & ELLIOTT, C. M., (1998), *Finite element error bounds for a curve shrinking with prescribed normal contact to a fixed boundary*, IMA J. Num. Anal., **18**, 635–654.
6. DECKELNICK, K., DZIUK, G. & ELLIOTT, C. M., (2005), *Computation of geometric partial differential equations and mean curvature flow*, Acta Numerica, **14**, 1–94.
7. DZIUK, G., (1991), *An algorithm for evolutionary surfaces* Numer. Math., **58**, 603–611.
8. DZIUK, G. & ELLIOTT, C.M., (2007), *Finite Elements on Evolving Surfaces*, IMA J. Num. Anal., **27**, 262–292.
9. DZIUK, G. & ELLIOTT, C.M., (2010), *An Eulerian approach to transport and diffusion on evolving implicit surfaces*, Computing and Visualization in Science, **13**, 17–28.
10. ELLIOTT, C.M., STINNER, B., (2009), *Analysis of a diffuse interface approach to an advection diffusion equation on a moving surface*, Math. Mod. Meth. Appl. Sci., **19**, 787–802.
11. ELLIOTT, C.M., STINNER, B., STYLES V. & WELFORD, R., (2011), *Numerical computation of advection and diffusion on evolving diffuse interfaces*, IMA J. Num. Anal., **31**, 786-812.
12. FIFE, P., CAHN, J. W. & ELLIOTT, C. M., (2001), *A free boundary model for diffusion-induced grain-boundary motion*, Interfaces and Free Boundaries, **3**, 291-336.
13. HANDWERKER, C., (1988), *Diffusion–induced grain boundary migration in thin films*, in Diffusion Phenomena in Thin Films and Microelectronic Materials, Ed. D. Gupta and P.S. Ho, Noyes Pubs. Park Ridge, N.J., 245–322.
14. MAYER, U.F. & SIMONETT, G., (1999), *Classical solutions for diffusion induced grain boundary boundary motion*, J. Math. Anal. Appl., **234**, 660–674.
15. RATZ, A. & VOIGT, A., (2006),*PDE's on surfaces - A diffuse interface approach*, Comm. Math. Sciences, **4**, 575–590.
16. SCHMIDT, A. & SIEBERT, K. G., (2005), *Design of adaptive finite element software: The finite element toolbox ALBERTA*, vol. 42 of Lecture notes in computational science and engineering, Springer.

Numerical Modeling of Stably Stratified Fluid Flow

L. Beneš, T. Bodnár, and J. Fürst

Abstract The article deals with the 2D numerical simulation of the stratified incompressible flows behind the moving thin horizontal strip in the towing tank and over the sinusoidal hill. The mathematical model is based on the Boussinesq approximation of the Navier–Stokes equations. The resulting set of PDE's is then solved by two different numerical methods. Different boundary conditions are tested.

1 Introduction

The influence of the stratification is significant in many processes taking place in ABL, sea, industry (e.g. stratification affects the transport of pollutants, plays significant role in determining the consequences of accidents on environment and human etc.). The simulations of stratified fluid flows are in general more demanding then the solution of similar non-stratified cases. Stratified flows is characterized by the variation of fluid density in the vertical direction that can lead to appearance of specific phenomena which are not present when density is constant, namely internal and gravity waves that propagate at long distances, jet–like flow structures, thin interfaces with high density and velocity gradients and anisotropic turbulence. The internal waves are particularly interesting since they effectively transport momentum but not mass. More about this type of flow can be found in e.g. [4, 5, 8]. Generation of the internal waves by moving body and behind the hill is studied numerically in this contribution.

L. Beneš (✉) · T. Bodnár · J. Fürst
Faculty of Mechanical Engineering, Department of Technical Mathematics, Czech Technical University in Prague, Karlovo náměstí 13, CZ-12135, Praha 2, Czech Republic
e-mail: benes@marian.fsik.cvut.cz; Tomas.Bodnar@fs.cvut.cz; Jiri.Furst@fs.cvut.cz

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_51,

2 Boussinesq Approximation

The flow is assumed to be incompressible, yet the density is not constant due to gravity. The set of Navier-Stokes equations for viscous incompressible flow with variable density was chosen as the mathematical model of this type of flow. These equations are simplified by the Boussinesq approximation. Full derivation of the basic equations is written in [3]. Resulting set of equations takes form

$$\frac{\partial \varrho}{\partial t} + \frac{\partial(\varrho u_j)}{\partial x_j} = -u_2 \frac{\partial \varrho_0}{\partial x_2}, \tag{1}$$

$$\frac{\partial u_i}{\partial t} + \frac{\partial(u_j u_i)}{\partial x_j} + \frac{1}{\varrho_*}\frac{\partial p}{\partial x_i} = \nu \frac{\partial^2 u_i}{\partial x_j\, \partial x_j} - \delta_{i,2} g, \tag{2}$$

$$\frac{\partial u_j}{\partial x_j} = 0, \tag{3}$$

where $W = [\varrho, u_1, u_2, p]^T$ is the vector of unknowns, $\varrho(x_1, x_2, t)$ denotes the perturbation of the density, ϱ_0 background density profile and u_1, u_2 are two velocity components, p stands for the pressure perturbation and g for the gravity acceleration. The x_1–axis is orientated in the direction of the motion and the x_2–axis is perpendicular to the density gradient.

The dimensionless parameters describing viscous stratified flow are defined by following relations. For the description of the stratified flows with characteristic velocity U and characteristic length L following parameters have been used:

$$Re = \frac{UL}{\nu}, \qquad Ri = -\frac{g}{\varrho_*}\frac{\frac{\partial \varrho_0}{\partial x_2}}{U}, \tag{4}$$

where U and L are the characteristic velocity and characteristic length, ν kinematic viscosity.

3 Numerical Schemes

Two different numerical schemes have been used for solution of the mentioned problems. The first scheme is the AUSM MUSCL scheme in the finite volume formulation combined with the artificial compressibility method, the second scheme is the compact finite–difference scheme.

- **AUSM scheme**

The first scheme is based on the artificial compressibility method in dual time. The continuity equation (3) is rewritten in the form

$$\frac{\partial p}{\partial \tau} + \beta^2 \frac{\partial u_j}{\partial x_j} = 0, \tag{5}$$

where τ is the artificial time. The finite volume AUSM scheme is used for the spatial semi-discretization of the inviscid fluxes. Quantities on the cell faces are computed using the MUSCL reconstruction with the Hemker-Koren limiter. The scheme is stabilized according to [6] by the pressure diffusion. The viscous fluxes are discretized using central approach on a dual mesh (diamond type scheme).

The spatial discretization results in a system of ODE's solved by the second-order BDF formula

$$\frac{3W^{n+1} - 4W^n + W^{n-1}}{2\Delta t} + L^{n+1} = 0. \tag{6}$$

Here, L^{n+1} denotes the numerical approximation of the convective and viscous fluxes described above and the source terms. Arising set of nonlinear equations is then solved by the artificial compressibility method in the dual time τ by the explicit three-stage second-order Runge-Kutta method.

- **Compact finite–difference scheme**

The second scheme is also based on the artificial compressibility method in dual time. The modified continuity equation (5) is used. The spatial semi-discretization is directly based on the paper [7], where the class of very high order compact finite difference schemes was introduced and analyzed. The main idea used to construct this family of schemes is that instead of approximating the spatial derivatives ϕ' of certain quantity ϕ explicitly from the neighboring values ϕ_i, the (symmetric) linear combination of neighboring derivatives $(\ldots, \phi'_{i-1}, \phi'_i, \phi'_{i+1}, \ldots)$ is approximated by weighted average of central differences. These schemes form a subclass of three-diagonal schemes with five-point computational stencil. The low-pass filter (for the filtered values $\widehat{\phi}_i$) of the following form was used:

$$b\,\widehat{\phi}_{i-1} + \widehat{\phi}_i + b\,\widehat{\phi}_{i+1} = 2\beta_0\phi_i + \beta_1 \frac{\phi_{i+1} + \phi_{i-1}}{2h} + \beta_2 \frac{\phi_{i+2} + \phi_{i-2}}{4h} + \ldots \tag{7}$$

Resulting system of ODE's has been solved by the three stage second order Strong Stability Preserving Runge-Kutta methods.

Both schemes were validated in our previous studies. The schemes have been successfully used for simulation of the flow field around moving bodies in 2D and 3D stratified fluid see [1, 2] and also for simulation of the stratified flow over the sinusoidal hill in 2D see [3] for wide range of Richardson numbers.

4 Computational Setup

The first problem solved in this study is the towing tank problem with moving thin horizontal strip 0.002×0.025 m. The case with the vertically placed strip was studied and validated in our previous studies [2]. The computational domain has dimensions

0.25×0.25 m. The strip is located in $x \in < 0.075, 0.1 >$ m and at the mid–heights. At the time $t = 0$ the obstacle starts moving to the right (in the positive x_1 direction) with constant velocity $U^{ob} = 0.0017$ m/s. The flow field is initially at rest with the exponential profile of stratification $\varrho_0 = \varrho_{00} \exp \frac{x_2}{\Lambda}$, $\varrho_{00} = 1{,}008.9\,\mathrm{kg/m^3}$, $\Lambda = 47.735$ m, the kinematic viscosity is $\nu = 10^{-6}\,\mathrm{m^2/s}$. This computational setup corresponds to the experimental setup of [4].

Homogeneous Neumann boundary conditions for all computed quantities were used in our computations. Fine Cartesian grid with 250×500 cells was used. It corresponds to resolution 1 mm in x_1 direction and 0.5 mm in x_2 direction. The obstacle is modeled by penalization technique as the source term in the momentum equations.

The second computational case is a half space with low smooth sine-shaped hill on the bottom. The whole domain has dimensions 90×30 m and the hill height is $h = 1$ m. The top of the hill is placed over the origin of the coordinate system.

The background density field is given by linear profile $\varrho_0(x_2) = \varrho_w + \gamma x_2$ with $\varrho_w = 1.2\,\mathrm{kg \cdot m^{-3}}$ and $\gamma = -0.01\,\mathrm{kg \cdot m^{-4}}$, the viscosity $\nu = 0.001$. Similar case was solved for wide range of Richardson numbers in [3]. In the presented study, the influence of the outlet boundary conditions on the generation of the gravity waves was tested. For the numerical simulations the case with $g = -50\,\mathrm{m\,s^{-2}}$ which corresponds to the Richardson number $Ri = 0.5$ was used.

The following boundary conditions are satisfied. On the inlet the velocity profile is given by the relation $u_1(x_2) = U_0(x_2/H)^{1/r}$ where $U_0 = 1$ m/s and $r = 40$ was prescribed, $u_2 = 0$, $\rho = 0$ and pressure perturbation is extrapolated. Homogeneous Neumann conditions are satisfied on the top. No–slip boundary conditions for the velocity components are prescribed and homogeneous boundary conditions for pressure and density perturbations are prescribed on the ground. The three different boundary conditions are prescribed on the outlet.

- **BC1:** homogeneous Neumann condition are prescribed for velocity components and pressure perturbations while pressure is set zero (homogeneous Dirichlet b.c.).
- **BC2:** advection equation $\frac{\partial q}{\partial t} + U_a \frac{\partial q}{\partial x} = 0$ is satisfied for $q = u_1, u_2, \varrho$, pressure is set zero. The advection velocity U_a is computed as the mean value of the u_1 velocity component on the inlet.
- **BC3:** are similar to previous case, only pressure is extrapolated.

The computations have been performed on structured non–orthogonal grid. The grid consists of 233×117 points refined near the ground and in the vicinity of the hill. The minimal resolution in the x_2 direction is $\Delta x_2 = 0.03$ m.

5 Numerical Results

Figure 1 shows the process of the wave generation in stably stratified flow. The flow pattern is typical for transient internal waves past an impulsively started body. The thin strip generates an initial perturbation and then gravity waves are formed.

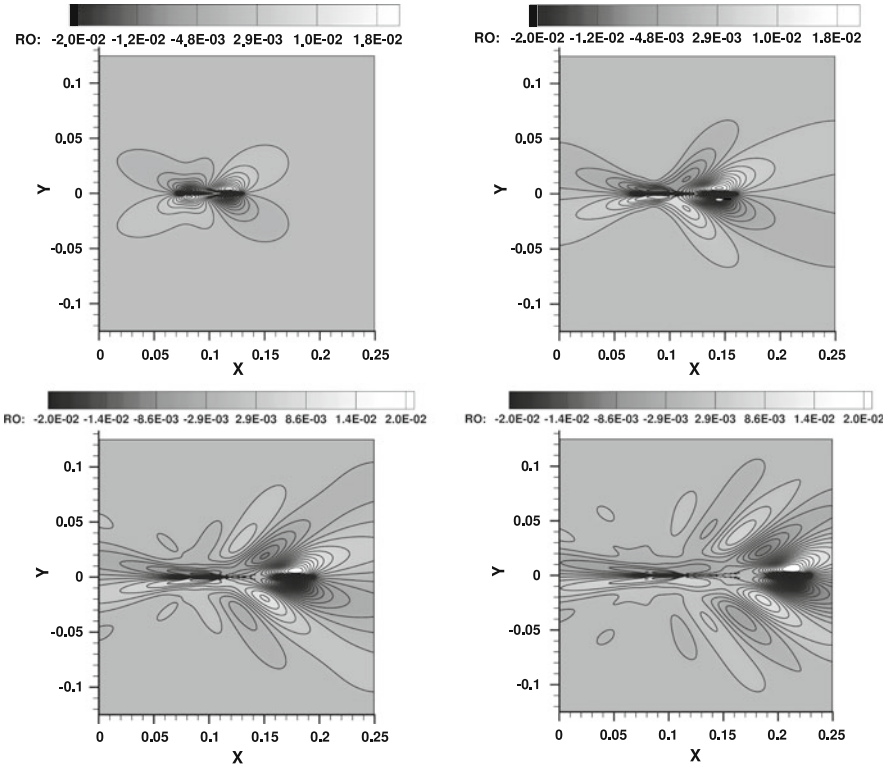


Fig. 1 Evolution of the density perturbance ϱ for four different times $t = 17.5\,\mathrm{s}$, $t = 35\,\mathrm{s}$, $t = 55\,\mathrm{s}$ and $75\,\mathrm{s}$. AUSM MUSCL scheme

The upstream disturbances are pronounced, what is typical for the flow with relatively low Froude number.

In Fig. 2, evolution of the profiles of the density perturbation ϱ in y-direction for two different position $x = 0.1\,\mathrm{m}$ and $x = 0.24\,\mathrm{m}$ is presented. Behind the obstacle strips with step–like density profile is formed (Fig. 2 – left). System of the gravity waves with wavelength given by the Brunt–Väisälä frequency forms in front of the obstacle.

Figure 3 displays the dependence of the flow on the boundary conditions. A comparison of the isolines of the u_2–velocity component for BC1–BC3 is presented at the same time. The gravity waves with the wavelength given by the Brunt–Väisälä frequencies are visible. The presented simulations are affected by some artifacts related to the implementation of the boundary conditions on the artificial boundaries of the computational domain. The wave pattern located close to the lower left corner of the domain is the most pronounced. This is a physical effect caused by the interaction of inlet flow profile with ground. It is a local effect in the non–stratified case. In stratified one this perturbation generates the second

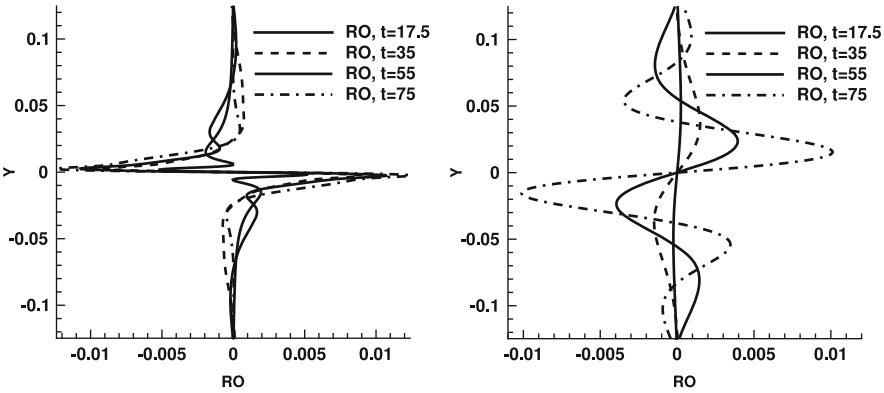


Fig. 2 Evolution of the ϱ profiles in y-direction for $x = 0.1$ m (*left*) and $x = 0.24$ m (*right*)

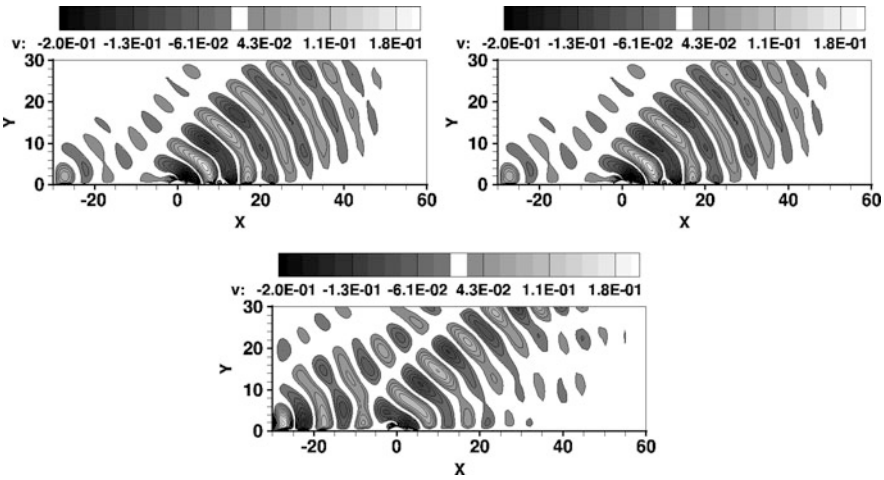


Fig. 3 Gravity waves pattern for three different boundary conditions. BC1 – *top left*, BC2 – *top right*, BC3 – *bottom*

system of gravity waves. This effect is strongest in the case BC3. The BC1 and BC2 produce very similar results. The effect of the corner waves is lover. The values of u_2 ranges for BC1 and BC2 in $u_2 \in < -0.228, 0.206 >$, for BC3 in $u_2 \in < -0.154, 0.207 >$. Similarly also density perturbations are lover in BC3 case ($\varrho \in < -0.0065, 0.0165 >$ BC1 and BC2, $\varrho \in < -0.0044, 0.0120 >$ BC3). It is also good visible in Figs. 4 and 5, where the dependence of u_1 and u_2 in two different positions $x = 0$ (top of the hill) and $x = 52$ (end of the domain) are shown. The profiles for BC1 and BC2 are practically the same. The wavelength is given by the Brunt–Väisälä frequency and is the same for all BC. The boundary layer in the case BC1 and BC2 is a little thinner with higher maximum on the top of the hill.

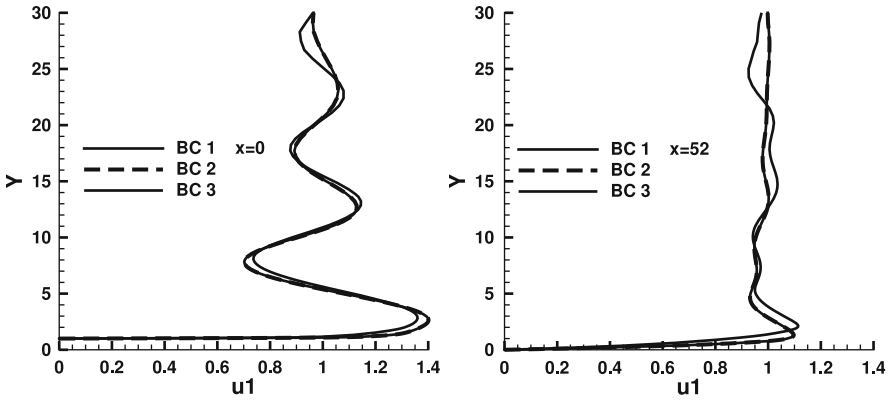


Fig. 4 Profiles of u_1 velocity component in y-direction for $x = 0\,\text{m}$ (*left*) and $x = 52\,\text{m}$ (*right*)

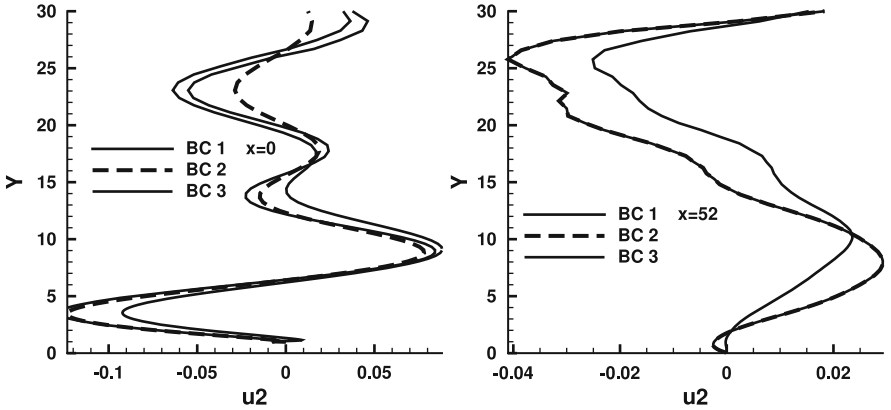


Fig. 5 Profiles of u_2 velocity component in y-direction for $x = 0\,\text{m}$ (*left*) and $x = 52\,\text{m}$ (*right*)

6 Conclusion

Two numerical schemes for stratified flows have been developed and they have been successfully used for simulation of the stratified flow around the moving body and over sinusoidal hill. Several numerical results have been obtained. The influence of different boundary conditions has been demonstrated. In the future the work will continue in the following directions: the detailed study of the flow structures in the internal and external aerodynamics and mainly numerical simulations of the flow over real orography.

Acknowledgements This work was supported by Research Plans MSM 6840770003, TACR Project TA01020428.

References

1. Beneš L., Fürst J., Fraunié Ph.: Numerical simulation of the stratified flow using high order schemes. *Engineering Mechanics*, 16(1):39–48, 2009. ISSN 1210-2717.
2. Beneš L., Fürst J., Fraunié Ph.: Comparison of two numerical methods for the stratified flow. *J. Computers & Fluids*,Vol:46(1) 2011 p. 148-154, ISSN 0045-7930.
3. Bodnár T., Beneš L., Fraunié Ph., Kozel K.: Application of compact finite–difference schemes to simulations of stably stratified fluid flows. *Applied Mathematics and Computation*, in press, DOI:10.1016/j.amc.2011.08.058.
4. Chaschechkin Y.D., Mitkin V.V.: Experimental study of a fine structure of 2D wakes and mixing past an obstacle in a continuously stratified fluid. *Dyn. Atmos. Oceans* 34, 165–187, 2001
5. Ding L., Calhoun R.J., Street R.L.: Numerical simulation of strongly stratified flow over three–dimensional hill. *Boundary–layer meteorology* 107: 81–114, 2003, Kluwer.
6. Dick E., Vierendeels J., Riemslagh K.: A multigrid semi–implicit line–method for viscous incompressible and low–mach–number flows on high aspects ratio grids. Journal of Computational Physics 154 310–341 (1999)
7. Lele S.K.: Compact finite difference schemes with spectral–like resolutions. *Journal of Computational Physics* 103(1992) 16–42.
8. Uchida T., Ohya Y.: A numerical study of stably stratified flows over a two dimensional hill – Part I, free–slip boundary conditions on the ground. *Journal of Wind Engineering and Industrial Aerodynamics* 67&68: 493–506, 1997.

Numerical Simulation of a Rising Bubble in Viscoelastic Fluids

H. Damanik, A. Ouazzi, and S. Turek

Abstract In this paper we discuss simulation techniques for a rising bubble in viscoelastic fluids via numerical methods based on high order FEM. A level set approach based on the work in (Sethian, Level set methods and fast marching methods: evolving interfaces in computational geometry, fluid mechanics, computer vision, and material science, 2nd edn. Cambridge University Press, 1999) is used for interface tracking between the bubble and the surrounding fluid. The two matters obey the Newtonian and the Oldroyd-B constitutive law in the case of a viscoelastic fluid while the flow model is given by the Navier-Stokes equations. The total system of equations is discretized in space by the LBB-stable finite element Q_2P_1, and in time by the family of θ-scheme integrators. The solver is based on Newton-multigrid techniques (Damanik et al., J Comput Phys 228:3869–3881, 2009; J Non-Newton Fluid Mech 165:1105–1113, 2010) for nonlinear fluids. First, we validate the multiphase flow results with respect to the benchmark results in (Hysing et al., Int J Numer Methods Fluids 60(11):1259–1288, 2009), then we perform numerical simulations of a bubble rising in a viscoelastic fluid and show cusp formation at the trailing edge.

1 Introduction

A rising bubble in a viscoelastic fluid shows a different behaviour than in a Newtonian fluid. In the first case, a cusp shape may appear at the trailing edge of the bubble while in the latter, this phenomena does not commonly appear. Thus, it is not only physically interesting but also numerically challenging to simulate such phenomena due to the multiphase characteristics and nonlinearity of the underlying

H. Damanik (✉) · A. Ouazzi, · S. Turek
Institut fuer Angewandte Mathematik, TU Dortmund, Vogelpothsweg 87, Dortmund, Germany
e-mail: hdamanik@mathematik.tu-dortmund.de;
Abderrahim.Ouazzi@mathematik.tu-dortmund.de; Stefan.Turek@mathematik.tu-dortmund.de

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_52,

fluid models. As already studied in the viscoelastic benchmark of flow around cylinder [2], the nonzero normal stress difference may exponentially increase behind a stagnation point of the cylinder to balance the almost zero shear rate. Similarly, the same mechanism can be used to describe the negative wake behind the bubble and leads to a cusp shape formation [6]. Experimental work has been done in [6] for different types of test liquids, size of the apparatus and volume of the bubbles which leads to a critical non-dimensional capillary number for cusp formation and velocity jump.

The idea to use a numerical approach to reproduce a cusp shape phenomena is not new. There are already methods for solving such complex systems such as lattice Boltzmann [10], finite element [3], finite difference schemes with boundary fitted orthogonal curvilinear coordinate systems [7] which are able to obtain qualitatively experimental results. Basically one needs both Newtonian and viscoelastic fluid models coupled with the well-known Navier-Stokes equations to describe the full domain system. Additionally, there is an interface tracking, resp., capturing algorithm one has to consider. Thus, the complete system of equations is very complex because the interface itself is part of the unknowns. Several tracking algorithms exist in the literature such as volume of fluid, phase field, level set method or coupled level set and volume of fluid [5]. In this study, the level set method of [9] is utilized without special objectivity. In the level set (ϕ) equation, the interface Γ_i between the two immiscible fluids changes its position with time (t) due to convection of the velocity ($\mathbf{u}$) only

$$\frac{\partial \phi}{\partial t} + (\mathbf{u} \cdot \nabla)\phi = 0. \tag{1}$$

The above Eq. (1) is a pure transport problem which determines the interface of the two (nonlinear) fluids. The good news is that the Newton-multigrid solver together with the high order finite element pair Q_2P_1 has proved successful with respect to such nonlinearity [1, 2]. The bad news is that solving the equation changes the property of the level set which should always be a distance function at every time step, that means

$$\|\nabla \phi\| = 1, \quad \forall \ t \in [0, T]. \tag{2}$$

The standard remedy is re-initialization. A brief overview of types of re-initialization can be seen in [3], see also the work of [8].

The objective of this study is to extend the application of our proposed numerical approach in [2] onto multiphase viscoelastic flow problems. Although a strong coupling of solving the discrete nonlinear system is being focused here, the level set equation is solved separately from the fluid flow model. The viscoelastic flow model is still solved in a fully coupled manner based on the previous work [1, 2]. In the next section, we start with the governing equations which then will be followed by the numerical solution technique section. Once we have validated the numerical approach with respect to a benchmark problem for a rising bubble in a Newtonian fluid [4], we describe numerical experiments of a rising bubble in viscoelastic fluids in the next section. We close the article with a summary in the last section.

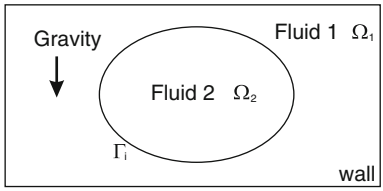


Fig. 1 General setup of two-phase problem with gravity. It shows a bubble placed in another fluid within rectangular geometry

2 Governing Equations

Fluid 1 is a viscoelastic fluid described by the Oldroyd-B model based on the conformation tensor τ (see Fig. 1),

$$\frac{\partial \tau}{\partial t} + \overbrace{(\mathbf{u}\cdot\nabla)\tau}^{\text{convection}} \underbrace{-\nabla\mathbf{u}\cdot\tau - \tau\cdot\nabla\,\mathbf{u}^T}_{\text{stretching}} + \frac{1}{\Lambda}(\tau - \mathbf{I}) = 0. \tag{3}$$

Fluid 2 is a liquid bubble given by a Newtonian model, $\tau_s = 2\eta_s(\mathbf{x},t)\mathbf{D}$ with constant viscosity $\eta_s(\mathbf{x},t)$ in each fluid phase. Here, $\mathbf{D} = \frac{1}{2}(\nabla\mathbf{u} + \nabla\mathbf{u}^{\mathrm{T}})$ is the deformation tensor. Both fluids are governed by the Navier-Stokes flow equations, i.e.

$$\begin{cases} \rho(\mathbf{x},t)\dfrac{\partial \mathbf{u}}{\partial t} + \rho(\mathbf{x},t)(\,\mathbf{u}\cdot\nabla)\mathbf{u} = \nabla\cdot\mathbf{T} + \rho(\mathbf{x},t)\mathbf{g} \\ \nabla\cdot\mathbf{u} = 0 \end{cases} \tag{4}$$

where the density depends on the position. The hydrostatic pressure p, the viscous- and the elastic-tensor contribute to the total stress tensor so that we can write

$$\mathbf{T} = -p\mathbf{I} + \tau_s + \frac{\eta_p(\,\mathbf{x},t)}{\Lambda}(\tau - \mathbf{I}). \tag{5}$$

In Fluid 2 the relaxation time goes to zero, $\Lambda = 0$, which from Eq. (3) it implies that the conformation stress tensor is equal to unity, $\tau = \mathbf{I}$. This condition defines a stress-less fluid (in the normal direction). Thus the last term of Eq. (5) cancels out.

As already mentioned, the material properties are constant in each fluid phase. Unfortunately since the interface is unknown one can not a priori set it to Dirichlet data as part of the discrete domain (mesh). Thus material properties change accordingly in particular to the position inside some thickness ϵ near the interface:

$$\rho(\mathbf{x},t) = \begin{cases} \rho_1, \ \forall\ \mathbf{x}\in\Omega_1(t) \\ \rho_2, \ \forall\ \mathbf{x}\in\Omega_2(t) \\ \rho_2 + \frac{1}{2}\left(1+\epsilon+\sin(\pi\epsilon)/\pi\right)(\rho_1-\rho_2), \ \forall\ \mathbf{x}\in\Gamma_i \end{cases} \tag{6}$$

The same rule is applied to the viscosity, $\eta_s(\mathbf{x},t)$ and $\eta_p(\,\mathbf{x},t)$.

3 Numerical Methods

We discretize the above system of equations with second order time integrators such as Crank-Nicolson which belongs to the family of θ-schemes. Given $\mathbf{u}^n$, ϕ^n, τ^n, $\rho(\mathbf{x}, t_n)$, $\eta_s(\mathbf{x}, t_n)$, $\eta_p(\mathbf{x}, t_n)$ and $\Delta t = t_{n+1} - t_n$, the first numerical step is to seek solutions $\mathbf{u}$, p, τ for the next time step

$$\rho(\mathbf{x}, t_n)\frac{\mathbf{u} - \mathbf{u}^n}{\Delta t} + \theta\left[\rho(\mathbf{x}, t_n)(\mathbf{u}\cdot\nabla\mathbf{u} - \mathbf{g}) - \eta_s(\mathbf{x}, t_n)\Delta\mathbf{u} - \frac{\eta_p(\mathbf{x}, t_n)}{\Lambda}\nabla\cdot\tau\right] + \nabla p$$

$$+ (1-\theta)\left[\rho(\mathbf{x}, t_n)(\mathbf{u}^n\cdot\nabla\mathbf{u}^n - \mathbf{g}) - \eta_s(\mathbf{x}, t_n)\Delta\mathbf{u}^n - \frac{\eta_p(\mathbf{x}, t_n)}{\Lambda}\nabla\cdot\tau^n\right] = 0 \quad (7)$$

$$\nabla\cdot\mathbf{u} = 0 \quad (8)$$

where $\mathbf{u}^n \sim \mathbf{u}(t_n)$. As one can see, the pressure space is discretized fully implicitly. The Oldroyd-B model is discretized simultaneously in the same way so that

$$\frac{\tau - \tau^n}{\Delta t} + \theta\left[\mathbf{u}\cdot\nabla\tau - \nabla\,\mathbf{u}\cdot\tau - \tau.\nabla\mathbf{u}^T + \frac{1}{\Lambda}(\tau - \mathbf{I})\right]$$

$$+ (1-\theta)\left[\mathbf{u}^n\cdot\nabla\tau^n - \nabla\,\mathbf{u}^n.\tau^n - \tau^n.\nabla\mathbf{u}^n + \frac{1}{\Lambda}(\tau^n - \mathbf{I})\right] = 0. \quad (9)$$

This maintains the monolithic character of solving and the accuracy of the solutions vector for the next numerical step. The second numerical step reads: Given $\phi^n, \mathbf{u}$, we seek solutions for the next time step of the level set equation

$$\frac{\phi - \phi^n}{\Delta t} + \theta\left[\mathbf{u}\cdot\nabla\phi\right] + (1-\theta)\left[\mathbf{u}\cdot\nabla\phi^n\right] = 0. \quad (10)$$

This step is exclusively decoupled from the first numerical step so that one can directly apply, if needed, standard numerical ingredients to stabilize the solver. In this step we observe that the high order finite element does not need numerical stabilization, but may not maintain mass conservation.

In each time step, the problem is discretized in space with the high order finite elements $Q_2P_1/Q_2/Q_2$ for velocity-pressure, (elastic) stress and level set.

The third numerical step is the re-initialization of ϕ. We recalculate the function value, to be almost exact, which is referred as "brute force", after finding all zero function values (interface). Thus, the accuracy depends on the number of interface points found on the finite element being used (see Fig. 2).

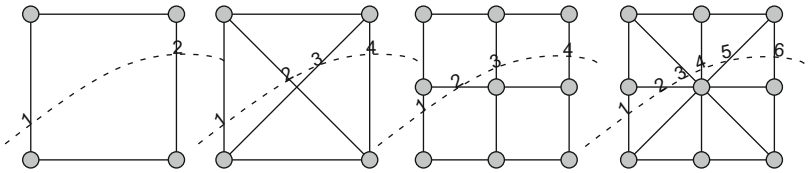


Fig. 2 Different possibilities of finding interface points in a finite element

4 Surface Tension

The surface tension of a liquid in real problem can not be neglected. One can consider it as additional volume force, $\mathbf{F}_{st}$, applied to the flow equation (4)

$$\mathbf{F}_{st} = \sigma\kappa\mathbf{n}. \tag{11}$$

The calculation of this force needs not only the gradient of level set $\mathbf{n}$ but also the curvature κ

$$\mathbf{n} = \nabla\phi, \quad \kappa = -\nabla\cdot\mathbf{n}. \tag{12}$$

While the gradient of the level set is provided within the finite element space, the curvature needs extra calculation after the third numerical step. One way to get this curvature is by projecting it into the finite element space $\tilde{\kappa} \in Q_2$ by solving

$$\int_\Omega \tilde{\kappa}\psi = -\int_\Omega (\nabla\cdot\mathbf{n})\psi \quad \forall\psi \in Q_2. \tag{13}$$

5 Rising Bubble Benchmark

5.1 Rising Bubble Benchmark in Newtonian Fluids

The classical benchmark of rising bubble is taken into account [4], please visit the webpage http://www.featflow.de/en/benchmarks/cfdbenchmarking/bubble.html for reference datas. Details of boundary conditions are also provided in [4]. We consider only test case 1 as comparison. A bubble is placed at the lower part of a 1×2 rectangular geometry with a radius of $r = 0.25$. Given a different density and viscosity between the two immiscible fluids, the bubble rises due to buoyancy force when solving the Navier-Stokes equations. Quantitative data measurements are undertaken in post-processing calculation, namely the center point of the bubble $\mathbf{X}_c = (x_c, y_c)$, the rising velocity $\mathbf{U}_c$ and the circularity ϕ of the bubble shape

$$\mathbf{X}_c = \frac{\int_{\Omega_2} \mathbf{x}\, dx}{\int_{\Omega_2} 1\, dx}, \quad \mathbf{U}_c = \frac{\int_{\Omega_2} \mathbf{u}\, dx}{\int_{\Omega_2} 1\, dx} \quad \phi = \frac{\pi d_a}{P_b} \tag{14}$$

where d_a is the diameter of area-equivalent circle and P_b is the perimeter of the bubble. The way to integrate these functional values may use the finite element function in each cubature point. A function value is included in the integration if the corresponding cubature point lies within the bubble domain Ω_2. In this way, it is efficient but it may not give accurate solution data. A more accurate way is to take the sum of all area of triangles and rectangles respectively within Ω_2.

5.2 Rising Bubble in Viscoelastic Fluids

There is no rigorous benchmark for a rising bubble in a viscoelastic liquid, but there exist numerical simulations for this purpose which tried to show cusp shape as observed in the experimental results. There are two types of simulations in literature: (i) A constant velocity data is given at the top boundary and then by buoyancy force the bubble remains still while a steady shape of the bubble can be obtained with marching of the time, see for example [7] and (ii) No velocity data is imposed and by buoyancy force the bubble rises and deforms its shape with time, i.e. [10]. Either way, the objective of the simulation is to obtain the cusp shape of the bubble. In this study we follow the second setup.

6 Numerical Results

6.1 Rising Bubble in Newtonian Fluids

The plots of circularity, center point, rising speed and the bubble shape are given in Fig. 3 which correspond to test case 1 in [4]. The high order finite element competes very fine with the ones from the benchmark that use low order FEM. The circularity of the bubble differs slightly at time $t = 3$. This can be due to no special treatment for mass loss when solving for the level set, while in the benchmark, artificial mass conservation as well as FEM-TVD stabilization has been used (by Group 1). However, the center point plot agrees very well with the reference data. The rising speed of the bubble agrees very well, too. Although the shape looses slightly mass at time $t = 3$, it can be accepted for the coarse mesh being used. We consider that our results are quite accurate with respect to the benchmark and continue with a rising bubble in viscoelastic fluids.

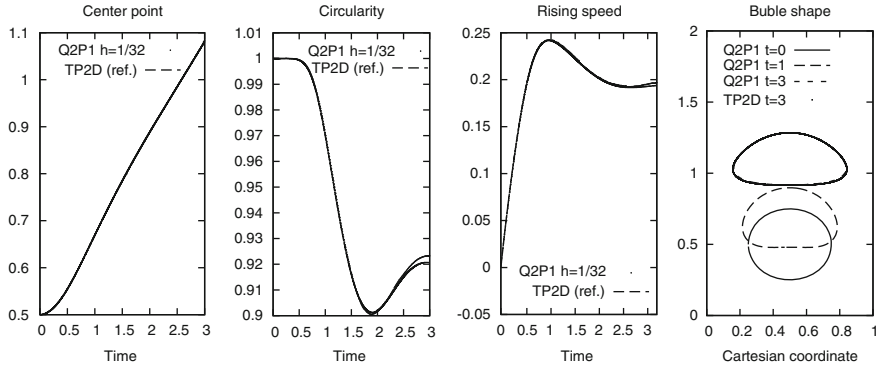


Fig. 3 Circularity and center point of the bubble the against benchmark data in [4]

6.2 Rising Bubble in Viscoelastic Fluids

The cusp formation is subject to certain conditions: The inertia effect (Re) must be small but the rising speed should be visible, the capillary number (Ca) should be bigger than some critical number and the Weissenberg number does not vanish (We $\neq 0$) [6]. Therefore the parameter setting from the previous benchmark is not suitable here. We re-set the parameter setting to be as in the following tabular to obtain different bubbles formation.

Test case	ρ_1	ρ_2	η_1	η_2	$\mathbf{g}$	σ
1. Viscoelastic ($\Lambda = 10$)	10	0.1	10	1	9.8	0.245
2. Newtonian ($\Lambda = 0$)	10	0.1	10	1	9.8	0.245
3. Viscoelastic ($\Lambda = 10$)	10	0.1	2	1	9.8	0.245
4. Newtonian ($\Lambda = 0$)	10	0.1	2	1	9.8	0.245

On the other hand the geometrical configuration is kept the same as in the previous benchmark because the shape of the column does not influence the cusp formation [6] as long as it is aligned with the gravity (the column does not tilt). The first two test cases (Fig. 4) show a different rising bubble behaviour in viscoelastic and Newtonian surrounding fluids. The cusp starts to appear very late at numerical time $t = 5$. On the other hand, the last two test cases (Fig. 4) show that the bubble rises faster than the first two test cases because of more inertia. Already at time $t = 3.4$ the bubble is close to the upper wall which shows cusp formation. Since the numerical time is shorter than the first two test cases, mass loss is less visible here.

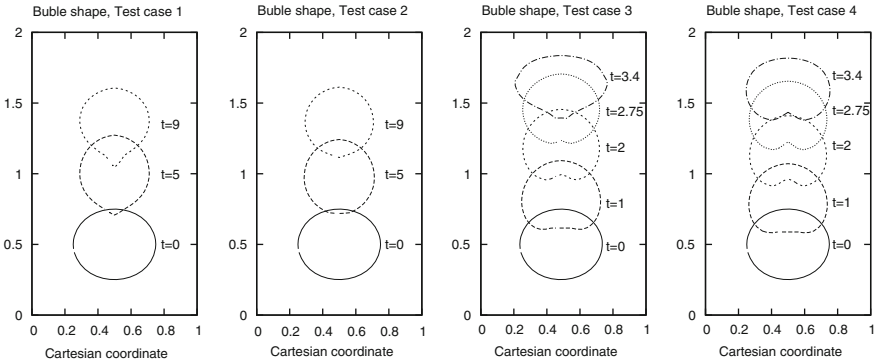


Fig. 4 Evolution of bubbles of test cases 1–4

7 Summary

We have analyzed numerical test cases that show cusp shape formation of a bubble when rising in viscoelastic fluids. A high order finite element approach is utilized for discretizing both fluid domains as well as the interface. The solver for the viscoelastic fluids is treated in a monolithic approach while the level set equation is solved separately from the fluid part. The numerical results are validated for a rising bubble in a Newtonian fluid which can be considered as quite accurate. Anyway, mass loss of the bubble is still present for the level set method with the chosen re-initialization. Further work is in progress in such a way that re-initialization does not introduce mass loss.

Acknowledgements The authors would like to thank the German Reasearch Foundation (DFG) for supporting the work through collaborative research center SFB/TR TRR 30 and SFB 708.

References

1. Damanik, H., Hron, J., Ouazzi, A., Turek, S.: A monolithic FEM–multigrid solver for non-isothermal incompressible flow on general meshes. Journal of Computational Physics **228**, 3869–3881 (2009)
2. Damanik, H., Hron, J., Ouazzi, A., Turek, S.: A monolithic FEM approach for the log-conformation reformulation (lcr) of viscoelastic flow problems. Journal of Non-Newtonian Fluid Mechanics **165**, 1105–1113 (2010)
3. Hysing, S.: Numerical simulation of immiscible fluids with FEM level set techniques. Ph.D. thesis, Technische Universität Dortmund (2007)
4. Hysing, S., Turek, S., Kuzmin, D., Parolini, N., Burman, E., Ganesan, S., Tobiska, L.: Quantitative benchmark computations of two-dimensional bubble dynamics. International Journal for Numerical Methods in Fluids **60 Issue 11**, 1259–1288 (2009)

5. Jimenez, E., Sussman, M., Otha, M.: A computational study of bubble motion in newtonian and viscoelastic fluids. Fluid Dynamics and Material Processing 1 **2**, 97–108 (2005)
6. Liu, Y. J., Liao, T. Y., Joseph, D. D.: A two-dimensional cusp at the trailing edge of an air buble rising in a viscoelastic liquid. Journal of Fluid Mechanics **304**, 321–342 (1995)
7. Noh, D. S., Kang, I. S., Leal, L. G.: Numerical solutions for the deformation of a buble rising in dilute polymeric fluids. Physics of Fluids **5**, 1315–1332 (1993)
8. Olsson, E., Kreiss, G.: A conservative level set method for two phase flow. Journal of Computational Physics **210**, 225–246 (2005)
9. Sethian, J.: Level Set Methods and Fast Marching Methods: Evolving Interfaces in Computational Geometry, Fluid Mechanics, Computer Vision, and Material Science. Cambridge University Press (1999). 2nd edition
10. Yoshino, M., Toriumi, Y., Arai, M.: Lattice boltzmann simulation of two-phase viscoelastic fluid flows. Journal of Computational Science and Technology **2**, 330–340 (2008)

A Reduced Model for Flow and Transport in Fractured Porous Media with Non-matching Grids

A. Fumagalli and A. Scotti

Abstract In this work we focus on a model reduction approach for the treatment of fractures in a porous medium, represented as interfaces embedded in a n-dimensional domain, in the form of a $(n-1)$-dimensional manifold, to describe fluid flow and transport in both domains. We employ a method that allows for non-matching grids, thus very advantageous if the position of the fractures is uncertain and multiple simulations are required. To this purpose we adopt an extended finite element approach, XFEM, to represent discontinuities of the variables at the interfaces, which can arbitrarily cut the elements of the grid. The method is applied to the solution of the Darcy and advection-diffusion problems in porous media.

1 Introduction

Subsurface flows are strongly influenced by the heterogeneities of the porous medium and in particular by the presence of fractures, faults and discontinuities between different layers. While micro-fractures can be accounted for by means of homogenization, large fractures and faults can act as preferential paths or barriers for the flow, and should be resolved by the grid. Since the characteristic width of these features is usually very small compared to the typical mesh size one possibility to address this problem in a computationally efficient way is to use a reduced model in which the fractures are represented as interfaces immersed in the porous medium, with proper coupling conditions between the fracture and the medium. The reduced model for the single phase Darcy problem was first introduced in [1] and extended in [2, 8]. In [4, 6] the authors extended the work of [8] allowing for non matching

A. Fumagalli (✉) · A. Scotti (✉)
MOX Laboratory, Department of Mathematics, Politecnico di Milano, Milan, Italy
e-mail: alessio.fumagalli@mail.polimi.it; anna.scotti@mail.polimi.it

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_53,

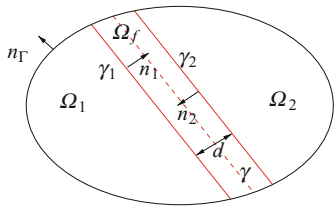


Fig. 1 Domain divided into two sub-domains Ω_1 and Ω_2 by a thin region Ω_f

grids between the porous domain and the fracture, increasing the flexibility of the method: an important advantage of non-matching grids is indeed the possibility to run multiple simulations with different fractures configurations without meshing each time the domain. In the present work we derive, in the same framework, a reduced model for the problem of the advection and diffusion of a tracer in a fractured porous medium, with the aim of providing a flexible and efficient tool to simulate realistic problems such as groundwater contamination. We obtain the advection field solving a Darcy problem, formulated as in [4], and employ the same space discretization with non-matching grids to approximate the transport problem.

The paper is structured as follows. In Sect. 2 the governing equations of Darcy flow and passive transport are presented. In Sect. 3 the reduced model for the transport problem is derived, and its numerical approximation is described in Sect. 4. In Sect. 5 two numerical tests are illustrated. Section 6 is devoted to conclusions.

2 Governing Equations

We consider the problem of a passive scalar, *e.g.* a tracer, transported by an external field $\boldsymbol{u}$ in a porous medium. In the case of our interest, $\boldsymbol{u}$ is obtained solving a Darcy problem. We are interested in the case of domains crossed by faults or large fractures characterized by a permeability tensor $\boldsymbol{K}$ that differs significantly from the porous matrix. From now on $i \in \{1, 2, f\}$ and $j \in \{1, 2\}$. Let us consider a regular domain $\Omega \in \mathbb{R}^n$, $n = 2$ or 3, with boundary $\overline{\Gamma} = \overline{\Gamma}_N \cup \overline{\Gamma}_D$ and outward unit normal $\boldsymbol{n}_\Gamma$, cut by a thin region $\Omega_f \subset \Omega$ of thickness d representing the fracture, as shown in Fig. 1, such that $\overline{\Omega} = \bigcup_i \overline{\Omega}_i$ and $\mathring{\Omega}_i \cap \mathring{\Omega}_j = \emptyset$ for $i \neq j$. The Darcy flow is described by

$$\begin{cases} \nabla \cdot \boldsymbol{u}_i = q_i \\ \boldsymbol{u}_i = -\boldsymbol{K}_i \left(\nabla p_i - \boldsymbol{q}_i \right) \end{cases} \text{ in } \Omega_i, \text{ with } \begin{cases} \boldsymbol{u}_j \cdot \boldsymbol{n}_j = \boldsymbol{u}_f \cdot \boldsymbol{n}_j \\ p_j = p_f \end{cases} \text{ on } \gamma_j, \tag{1}$$

where the subscript i denotes the quantity in each Ω_i and $\gamma_j \in \mathbb{R}^{n-1}$ is the interface between Ω_j and Ω_f with outward unit normal $\boldsymbol{n}_j$. We impose on (1) boundary conditions $p_i = \overline{p}_i$ on Γ_N^p and $\boldsymbol{u}_i \cdot \boldsymbol{n}_\Gamma = \overline{g}_i$ on Γ_D^p.

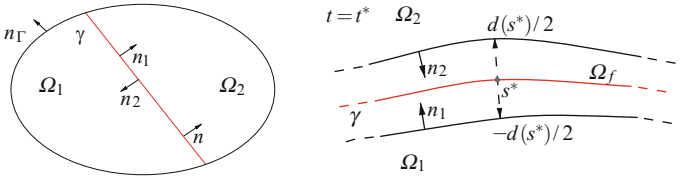


Fig. 2 *Left*: domain cut by an $1D$ interface γ that replaces Ω_f. *Right*: the reducing process

Moving to the advection-diffusion problem, we indicate with c the concentration of the passive scalar, defined as the volume fraction of tracer in the porosity, the total flux $\boldsymbol{\chi} := -\boldsymbol{D}\nabla c + \boldsymbol{u}c$ and we denote with $\boldsymbol{D} \in \left[L^\infty(\Omega)\right]^{n\times n}$ the molecular diffusion tensor, which is symmetric and positive definite. Introducing the interval of time $\mathscr{I}_T := (0, T)$ and the domain $Q_i := \Omega_i \times \mathscr{I}_T$ then the system of equations is

$$\begin{cases} \Phi_i \dfrac{\partial c_i}{\partial t} + \nabla\cdot \boldsymbol{\chi}_i = g \\ \boldsymbol{\chi}_i = -\boldsymbol{D}_i \nabla c_i + \boldsymbol{u}_i c_i \end{cases} \text{ in } Q_i, \text{ with } \begin{cases} \boldsymbol{\chi}_j \cdot \boldsymbol{n}_j = \boldsymbol{\chi}_f \cdot \boldsymbol{n}_j \\ c_j = c_f \end{cases} \text{ on } \gamma_j \times \mathscr{I}_T. \quad (2)$$

Here $\Phi_i \in L^\infty(\Omega)$ denotes the porosity and g a source term. We impose on (2) $c_i = \overline{c}$ on $\Gamma_N^c \times \mathscr{I}_T$, $\boldsymbol{\chi}_i \cdot \boldsymbol{n}_\Gamma = \overline{\chi}_i$ on $\Gamma_D^c \times \mathscr{I}_T$ and $c_i = c_0$ in $\Omega_i \times \{0\}$.

3 Reduced Model for the Advection-Diffusion Problem

We want to derive a reduced model for advection and diffusion in the presence of fractures, replacing Ω_f with a $n-1$ dimensional interface $\gamma \approx \gamma_j$ with unit normal $\boldsymbol{n} \approx \boldsymbol{n}_1 \approx -\boldsymbol{n}_2$, as shown in Fig. 2.

In [8] a reduced model for Darcy is derived, yielding two coupled problems for the flow in γ and in the porous matrix. We report the main results for readers convenience. Given a function $a : \Omega \to \mathbb{R}^m$, $m = 1$ or n, let us set $[\![a]\!]_\gamma := a_1 - a_2$ and $\{\{a\}\}_\gamma := (a_1 + a_2)/2$ with $a_j(\boldsymbol{x}) = \lim_{\epsilon\to 0^\pm} a(\boldsymbol{x} - \epsilon\boldsymbol{n})$. We define the projection matrix $\boldsymbol{N} := \boldsymbol{n} \otimes \boldsymbol{n}$. Indicating with $\hat{\cdot}$ the reduced variables defined in γ, following [8] we suppose $\boldsymbol{K}_f = K_{f,\boldsymbol{n}}\boldsymbol{N} + K_{f,\boldsymbol{\tau}}(\boldsymbol{I} - \boldsymbol{N})$, then (1) becomes

$$\begin{cases} \nabla\cdot \boldsymbol{u}_j = q_j \\ \boldsymbol{u}_j = -\boldsymbol{K}_j\left(\nabla p_j - \boldsymbol{q}_j\right) \end{cases} \text{ in } \Omega_j, \quad \begin{cases} \nabla_{\boldsymbol{\tau}}\cdot \hat{\boldsymbol{u}} = \hat{q} + [\![\boldsymbol{u}\cdot\boldsymbol{n}]\!]_\gamma \\ \hat{\eta}\hat{\boldsymbol{u}} + \nabla_{\boldsymbol{\tau}}\hat{p} = \hat{\boldsymbol{q}} \end{cases} \text{ in } \gamma, \quad (3)$$

where $\hat{\eta} := d/K_{f,\boldsymbol{\tau}}$ and $p_j = \overline{p}_j$ on Γ_N^p, $\boldsymbol{u}_j \cdot \boldsymbol{n}_\Gamma = \overline{g}_j$ on Γ_D^p, $\hat{p} = \hat{\overline{p}}$ on $\partial\gamma \cap \Gamma_N^p$ and $\hat{\boldsymbol{u}} \cdot \boldsymbol{n}_\Gamma = \hat{\overline{g}}$ on $\partial\gamma \cap \Gamma_D^p$. Furthermore given $a : \Omega \to \mathbb{R}$ and $\boldsymbol{a} : \Omega \to \mathbb{R}^n$, we

have $\nabla_{\boldsymbol{\tau}} a := \nabla a - \boldsymbol{N}\nabla a$ and $\nabla_{\boldsymbol{\tau}} \cdot \boldsymbol{a} := \nabla \cdot \boldsymbol{a} - \boldsymbol{N} : \nabla \boldsymbol{a}$. The coupling conditions, derived in [8] for $\boldsymbol{q} \equiv 0$, become in the more general case

$$\begin{cases} \xi_0 \eta_\gamma [\![\boldsymbol{u} \cdot \boldsymbol{n}]\!]_\gamma + \dfrac{d}{4} [\![\boldsymbol{q} \cdot \boldsymbol{n}]\!]_\gamma = \{\!\{ p \}\!\}_\gamma - \hat{p} \\ \eta_\gamma \{\{ \boldsymbol{u} \cdot \boldsymbol{n} \}\}_\gamma = [\![p]\!]_\gamma + d \{\!\{ \boldsymbol{q} \cdot \boldsymbol{n} \}\!\}_\gamma \end{cases} \quad \text{on } \gamma,$$

where $\xi_0 \in (0,\ 0.25]$ is a shape parameter and $\eta_\gamma := 1/\left(d K_{f,\boldsymbol{n}}\right)$, see [5] for details.

We derive, in an analogous way, a reduced model for (2). To this purpose we define the reduced flux and the mean concentration in γ, see Fig. 2, as

$$\hat{\boldsymbol{\chi}}(\boldsymbol{s}, t) := \int_{-\frac{d}{2}}^{\frac{d}{2}} \boldsymbol{\chi}_{f,\boldsymbol{\tau}}(\boldsymbol{r}, t)\, d\boldsymbol{r} \quad \text{and} \quad \hat{c}(\boldsymbol{s}, t) := \frac{1}{d} \int_{-\frac{d}{2}}^{\frac{d}{2}} c_f(\boldsymbol{r}, t)\, d\boldsymbol{r},$$

where $\boldsymbol{s} \in \gamma$, $d = d(\boldsymbol{s})$ and for a function $\boldsymbol{a}_f : \Omega_f \to \mathbb{R}^n$, $\boldsymbol{a}_{f,\boldsymbol{\tau}} := \boldsymbol{a}_f - \boldsymbol{N}\boldsymbol{a}_f$. Projecting the conservation equation on the tangential space of γ and integrating in each section of γ, the reduced conservation equation becomes

$$d\Phi_f \frac{\partial \hat{c}}{\partial t} + \nabla_{\boldsymbol{\tau}} \cdot \hat{\boldsymbol{\chi}} = \hat{g} + [\![\boldsymbol{\chi} \cdot \boldsymbol{n}]\!]_\gamma \quad \text{on } \gamma \times \mathscr{I}_T, \tag{4}$$

where $\hat{g}$ is the reduced scalar source. We have assumed that Φ_f is constant in each transversal section of γ. Projecting the second equation of (2) on the tangential space of γ and integrating in each section of γ, the reduced flux equation becomes

$$\hat{\beta}\hat{\boldsymbol{\chi}} + \nabla_{\boldsymbol{\tau}} \hat{c} - d\hat{\beta}\hat{\boldsymbol{u}}\hat{c} = \boldsymbol{0} \quad \text{on } \gamma \times \mathscr{I}_T, \tag{5}$$

where $\hat{\beta} := d/D_{f,\boldsymbol{\tau}}$ and with $\int_{-\frac{d}{2}}^{\frac{d}{2}} \boldsymbol{u}_{f,\boldsymbol{\tau}} c_f\, d\boldsymbol{r} \approx d\hat{\boldsymbol{u}}\hat{c}$ and $\int_{-\frac{d}{2}}^{\frac{d}{2}} \boldsymbol{u}_f \cdot \boldsymbol{n} c_f\, d\boldsymbol{r} \approx 0$. We then integrate the second equation in (2) along the normal direction in Ω_f, apply the trapezium quadrature rule and exploit the continuity on γ_1 and γ_2 to obtain

$$\beta_\gamma \{\{ \boldsymbol{\chi} \cdot \boldsymbol{n} \}\}_\gamma = [\![c]\!]_\gamma \quad \text{on } \gamma \times \mathscr{I}_T, \tag{6}$$

where $\beta_\gamma := 1/\left(d D_{f,\boldsymbol{n}}\right)$. To close the reduced system we need another relation to model the variation of the concentration and total flux across the fracture. From a Taylor expansion in the centre of Ω_f, see [5], we find

$$\hat{c} = \{\!\{ c \}\!\}_\gamma - \beta_\gamma \xi_0 [\![\boldsymbol{\chi} \cdot \boldsymbol{n}]\!]_\gamma \quad \text{on } \gamma \times \mathscr{I}_T, \tag{7}$$

where $\xi_0 \in (0,\ 0.25]$ accounts for different concentration profiles in Ω_f. Using (4) and (5) we write the system for the advection-diffusion problem as

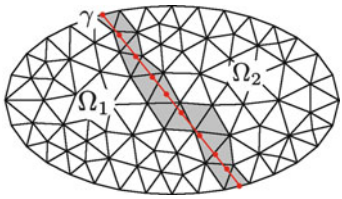


Fig. 3 The triangulation of Ω and γ. The mesh elements cut by γ are highlighted

$$\begin{cases} \Phi_j \dfrac{\partial c_j}{\partial t} + \nabla \cdot \boldsymbol{\chi}_j = g_j \\ \boldsymbol{\chi}_j = -\boldsymbol{D}_j \nabla c_j + \boldsymbol{u}_j c_j \end{cases} \text{in } Q_j, \quad \begin{cases} d\Phi_f \dfrac{\partial \hat{c}}{\partial t} + \nabla_{\boldsymbol{\tau}} \cdot \hat{\boldsymbol{\chi}} = \hat{g} + [\![\boldsymbol{\chi} \cdot \boldsymbol{n}]\!]_\gamma \\ \hat{\beta} \hat{\boldsymbol{\chi}} + \nabla_{\boldsymbol{\tau}} \hat{c} - d\hat{\beta} \hat{\boldsymbol{u}} \hat{c} = \boldsymbol{0} \end{cases} \text{in } \gamma \times \mathscr{I}_T, \tag{8}$$

coupled with the interface conditions (6) and (7) on γ and complemented with $c_j = \overline{c}_j$ on $\Gamma_N^c \times \mathscr{I}_T$, $\boldsymbol{\chi}_j \cdot \boldsymbol{n}_\Gamma = \overline{\chi}_j$ on $\Gamma_D^c \times \mathscr{I}_T$, $c_i = c_{0,i}$ in $\Omega_i \times \{0\}$. Further for γ we have $\hat{c} = \hat{\overline{c}}_f$ on $\partial\gamma_N^c \times \mathscr{I}_T$, $\hat{\boldsymbol{\chi}} \cdot \boldsymbol{n}_\Gamma = \hat{\overline{\chi}}_f$ on $\partial\gamma_D^c \times \mathscr{I}_T$ and $\hat{c} = \hat{c}_{0,f}$ in $\gamma \times \{0\}$.

4 Numerical Discretization

We choose to adopt the same space discretization for the Darcy problem and for the advection-diffusion problem. In particular, we employ the mixed finite element method with the lowest order Raviart-Thomas finite elements $\mathbb{RT}_0$ while time stepping is performed, for the time dependent problem, by the implicit Euler method, as in [10]. The same discretization strategy is employed for the transport problem in the bulk medium and in the fracture. Mixed finite element are a valuable choice in problems concerning flow in porous media thanks to their local mass conservation property. Moreover, they have been successfully applied to problems of transport in porous media, see [3, 10].

We are interested in the case of domains crossed by interfaces that are non-conforming with the grid. More precisely, the triangulation $\mathscr{T}_h$ of Ω and that of the interface γ are completely independent and non-matching, as shown in Fig. 3. To this purpose, we adopt an extended finite element (XFEM) approach [9], enriching the classical $\mathbb{RT}$ finite element basis on the elements cut by γ with discontinuous functions. In particular, we follow the approach proposed in [7], and proceed, for the transport problem in the fractured medium, as proposed in [4] for the Darcy problem. We consider discrete fluxes $\boldsymbol{\chi}_h \in \boldsymbol{V}_h$ and concentration $c_h \in Q_h$ made of two components, associated to Ω_j where $\boldsymbol{V}_h := \boldsymbol{V}_{1,h} \times \boldsymbol{V}_{2,h}$ and $Q_h := Q_{1,h} \times Q_{2,h}$, with

$$\boldsymbol{V}_{j,h} := \left\{ \boldsymbol{v}_h \in \boldsymbol{H}_{\text{div}}(\Omega_j) : \ \boldsymbol{v}_h|_{K_l} \in \mathbb{RT}_0(K_l), \ K_l \in \mathscr{T}_h \right\},$$

$$Q_{j,h} := \left\{ q_h \in L^2(\Omega_j) : \ q_h|_{K_l} \in \mathbb{P}_0(K_l), \ K_l \in \mathscr{T}_h \right\},$$

where, for any $K_l \in \mathcal{T}_h \cap \Omega_j$, $\mathbb{RT}_0(K_l)$ and $\mathbb{P}_0(K_l)$ are the restrictions to K_l of the standard $\mathbb{RT}_0$ and $\mathbb{P}_0$ local functions. The discrete variables can thus be discontinuous on γ, being defined on each part K_l of a cut element K by independent functions.

The global coupled system, discretized in space and time, reads

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B}^\top & \boldsymbol{0} & \boldsymbol{E} \\ \boldsymbol{B} & \boldsymbol{M} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \hat{\boldsymbol{A}} & \hat{\boldsymbol{B}}^\top \\ \boldsymbol{E}^\top & \boldsymbol{0} & \hat{\boldsymbol{B}} & \hat{\boldsymbol{M}} \end{bmatrix} \begin{bmatrix} \boldsymbol{\chi}_h \\ \boldsymbol{c}_h^{k+1} \\ \hat{\boldsymbol{\chi}}_h \\ \hat{\boldsymbol{c}}_h^{k+1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{0} \\ -\boldsymbol{g} + \boldsymbol{M}\boldsymbol{c}^k \\ \boldsymbol{0} \\ -\hat{\boldsymbol{g}} + \hat{\boldsymbol{M}}\hat{\boldsymbol{c}}^k \end{bmatrix}$$

where the blocks $\boldsymbol{E}$ and $\boldsymbol{E}^\top$ account for the coupling between the two problems and for the interpolation between the bulk mesh and the fracture mesh, furthermore $\boldsymbol{M}$ and $\hat{\boldsymbol{M}}$ are the mass matrices which include the time step Δt. We point out that in the relevant case of advection dominated problems a stabilization has to be applied, see [10, 11] for stabilization techniques in mixed finite elements.

5 Results

Test case 1

Let $\Omega = [0,1]^2$, $\Gamma = \{(x,y) \in \Omega : \ y = 2x - 0.4\}$, $\Gamma_D = \{0,1\} \times [0,1]$, and $\Gamma_N = [0,1] \times \{0,1\}$. The flow is described by (3), with $q = \hat{q} = 0$, $\boldsymbol{q} = \hat{\boldsymbol{q}} = 0$, $\overline{p} = 1 - y$, and $d = 0.01$. We consider full Neumann boundary conditions $\hat{p} = 1 - y$ on $\partial\gamma_N$, coupled with the interface conditions with $\xi_0 = 1/8$. The permeability tensor of the medium is isotropic, $\boldsymbol{k}_m = \boldsymbol{I}$ while γ is characterized by a high permeability in the normal and tangential directions, $\boldsymbol{k}_f = 100\boldsymbol{I}$. We solve (2) where the advection field is the computed Darcy velocity which is higher in γ than in the porous matrix. We set $\overline{c} = 0$, $\overline{\chi} = 0$ and $c_0(x,y) = 1$ if $(x - 0.5)^2 + (y - 0.2)^2 < 0.03$. The diffusion tensor is $\boldsymbol{D} = 0.05\boldsymbol{I}$. We first solve this test case with the standard mixed FEM and a refined mesh that is able to resolve the fracture and compare the results with the reduced model and the XFEM approach. The time step is $\Delta t = 5 \cdot 10^{-3}$ and $T = 0.2$. In Fig. 4 the solutions are compared at two different times. In both cases the tracer is advected upwards and flows preferably along the fracture where the fluid velocity is higher. The two methods produce results that are qualitatively in good agreement, even if a grid of only 3,200 triangles and 100 segments for the fracture is used with the XFEM method and a much more refined grid of ∼13,000 triangles is needed with the standard approach.

Test case 2

We present a realistic example. The domain, cut by γ, sketched in Fig. 5, has spatial dimensions $6Km \times 3Km$ while the end time is $T = 10^{13}$ s $\simeq 0.95$ Ma. The data for (3) and (8) are reported in Fig. 5. Note that the permeability is isotropic

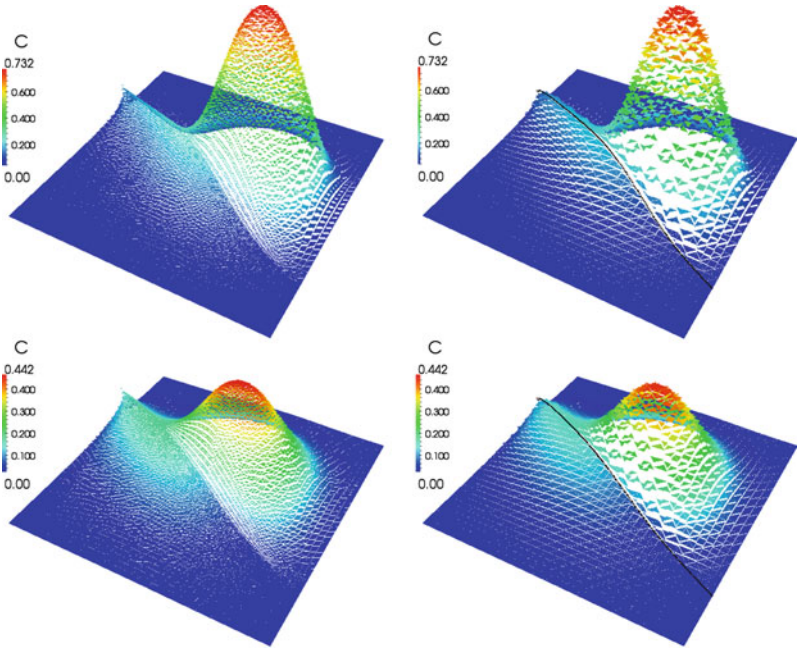
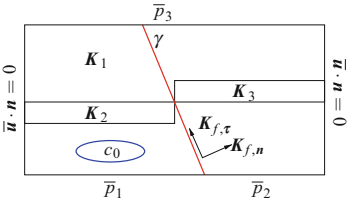


Fig. 4 On the *left* the tracer computed with the standard FEM and the fine grid at time $t = 0.1$ and $t = 0.2$, respectively. On the *right* the solution obtained at the same time with the reduced approach and a coarse grid. The *black line* represents the concentration along γ



$K_1 = 10^{-15} m^2$	$K_2 = 10^{-18} m^2$	$K_3 = 10^{-17} m^2$
$K_{f,\boldsymbol{\tau}} = 10^{-13} m^2$	$K_{f,\boldsymbol{n}} = 10^{-17} m^2$	$\overline{p}_1 = 34.53\, MPa$
$\overline{p}_2 = 29.53\, MPa$	$\overline{p}_3 = 0.1\, MPa$	$\boldsymbol{q} = -9.91 \boldsymbol{e}_2$
$d = 10 m$	$\xi_0 = 0.25$	$q = 0$
$\Delta t = 5 \cdot 10^{10} s$	$c_0 = 0.3$	$D = 10^{-8} m^2/s$

Fig. 5 *Left*: computational domain showing. *Right*: data for problems (3) and (8)

in the medium while γ acts as a preferential path for the flux in the tangential direction and as a barrier in the normal direction. The molecular diffusion of the tracer is homogeneous and isotropic in the whole domain, thus in γ we set $D_{f,\boldsymbol{n}} = Dd$ and $D_{f,\boldsymbol{\tau}} = D/d$. We impose homogeneous essential conditions on the left and the right part of Γ and natural conditions on the top and the bottom of Γ. Figure 6 shows the concentration of the tracer at two different times comparing the results obtained accounting for γ and neglecting it. We notice that the solutions are extremely different confirming the necessity to handle the fractures in an efficient and accurate way.

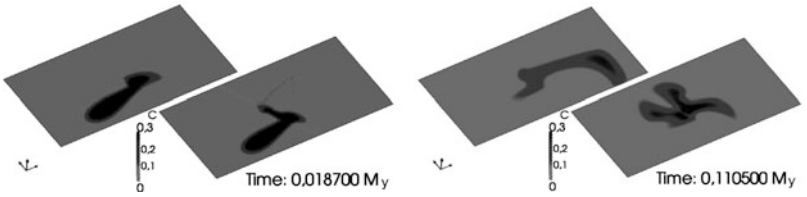


Fig. 6 Comparison, at two different times, between the solution obtained without the fracture, on the *top*, and the solution with the fracture and the reduced model, on the *bottom*

6 Conclusions

In this paper we presented an original model for transport problems in fractured porous media. Following the approach present in the literature for the single phase Darcy flow we derived a numerical model for the coupled problem of advection and diffusion in the porous medium and in the fractures, and compared the results with the traditional approach. Thanks to its moderate computational cost the method proves to be effective for cases with realistic parameters. Our future work will focus on the assessment of the theoretical properties of the method and the inclusion in this framework of suitable stabilization techniques for advection dominated problems.

References

1. Clarisse Alboin, Jérôme Jaffré, Jean E. Roberts, Xuewen Wang, and Christophe Serres. *Domain decomposition for some transmission problems in flow in porous media*, volume 552 of *Lecture Notes in Phys*., pages 22–34. Springer, Berlin, 2000.
2. Philippe Angot, Franck Boyer, and Florence Hubert. Asymptotic and numerical modelling of flows in fractured porous media. *M2AN Math. Model. Numer. Anal*., 43(2):239–275, 2009.
3. Fabian Brunner, Florin Adrian Radu, Markus Bause, and Peter Knabner. Optimal order convergence of a modified BDM_1 mixed finite element scheme for reactive transport in porous media. *Advances in Water Resources*, 35:163–171, 2012.
4. Carlo D'Angelo and Anna Scotti. A mixed finite element method for Darcy flow in fractured porous media with non-matching grids. *Mathematical Modelling and Numerical Analysis*, 46(02):465–489, 2012.
5. Alessio Fumagalli. *Numerical Modelling of Flows in Fractured Porous Media by the XFEM Method*. PhD thesis, Politecnico di Milano, 2012.
6. Alessio Fumagalli and Anna Scotti. Numerical modelling of multiphase subsurface flow in the presence of fractures. *Communications in Applied and Industrial Mathematics*, December 2011. In press.
7. Anita Hansbo and Peter Hansbo. An unfitted finite element method, based on Nitsche's method, for elliptic interface problems. *Comput. Methods Appl. Mech. Engrg*., 191(47–48):5537–5552, 2002.
8. Vincent Martin, Jérôme Jaffré, and Jean E. Roberts. Modeling fractures and barriers as interfaces for flow in porous media. *SIAM J. Sci. Comput*., 26(5):1667–1691, 2005.
9. Nicolas Moës, John Dolbow, and Ted Belytschko. A finite element method for crack growth without remeshing. *Int. J. for Numerical Methods in Eng*., 46(1):131–150, 1999.

10. Florin Adrian Radu, Iuliu Sorin Pop, and Sabine Attinger. Analysis of an Euler implicit - mixed finite element scheme for reactive solute transport in porous media. *Numer. Methods Part. Differ. Equat*., 26:320–344, 2010.
11. Riccardo Sacco and Fausto Saleri. Stabilized mixed finite volume methods for convection-diffusion problems. *East-West J. Numer. Math*., 5(4):291–311, 1997.

Higher Order Galerkin Time Discretization for Nonstationary Incompressible Flow

S. Hussain, F. Schieweck, and S. Turek

Abstract In this paper, we extend our work for the heat equation in (Hussain et al., J Numer Math 19(1):41–61, 2011) and for the Stokes equations in (Hussain et al., Open Numer Methods J 4:35–45, 2012) to the nonstationary Navier-Stokes equations in two dimensions. We examine *continuous* Galerkin-Petrov (cGP) time discretization schemes for nonstationary incompressible flow. In particular, we implement and analyze numerically the higher order cGP(2)-method. For the space discretization, we use the LBB-stable finite element pair Q_2/P_1^{disc}. The discretized systems of nonlinear equations are treated by using the fixed-point as well as the Newton method and the associated linear subproblems are solved by using a monolithic multigrid solver with GMRES method as smoother. We perform nonstationary simulations for a benchmarking configuration to analyze the temporal accuracy and efficiency of the presented time discretization scheme.

1 Introduction

A class of time discretization schemes which is based on Rothe's method is the *continuous* Galerkin-Petrov discretization (cGP(k)-methods) and *discontinuous* Galerkin (dG(k)-methods). The approach of the cGP-method has already been used by Aziz and Monk [1] (but not under this name) for the linear heat equation. These

S. Hussain (✉) · S. Turek
Institut für Angewandte Mathematik, TU-Dortmund, Vogelpothsweg 87, 44227, Dortmund, Germany
e-mail: shafqat.hussain@math.tu-dortmund.de; stefan.turek@math.tu-dortmund.de

F. Schieweck
Institut für Analysis und Numerik, Otto-von-Guericke-Universität Magdeburg, Postfach 4120, D-39016, Magdeburg, Germany
e-mail: schiewec@ovgu.de

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_54,

time discretizations are found to be of higher order and have been studied for the heat equation in [3] and for the Stokes equations in [5].

In this paper, we want to extend this numerical study for the nonstationary Navier-Stokes equations. In particular, we implement and analyze numerically the cGP(2)-method which is found to be of higher order at comparable numerical cost. The cGP(2)-method is of order 3 in the whole time interval and superconvergent of order 4 in the discrete time points. The spatial discretization is carried out by using biquadratic finite elements for the velocity and discontinuous linear elements for pressure. From the numerical studies [3,5], we have observed that the estimated experimental orders of convergence confirm the expected theoretical orders. Furthermore, the tests have shown that the cGP(2)-scheme provides significantly more accurate numerical solutions for both velocity and pressure than the other presented schemes cGP(1) and dG(1)-method (see [3,5] for comparison).

Since we obtain superconvergence results for the velocity only at the discrete time points t_n, it is also desirable to get a high order pressure at the same points, for instance, for the computation of the hydrodynamic forces in CFD problems such as drag, lift etc. In order to get a higher order pressure, we perform a special post processing as described in [5].

The resulting discretized system of nonlinear equations which is characterized as a saddle point problem is treated by using the fixed-point and Newton method. The associated linear subproblems are solved by using a coupled multigrid solver with a local pressure Schur complement type smoother. Finally, we perform simulations for nonstationary flow problems to demonstrate the high accuracy of the cGP(2)-method. The test problem which is considered in this paper corresponds to the classical 'flow around cylinder' benchmark [8].

2 Galerkin Time Stepping for the Navier-Stokes Equations

We consider the nonstationary incompressible Navier-Stokes equations, i.e. we want to find a velocity $\mathbf{u}$ and a pressure p *such that*

$$\begin{aligned} \partial_t \mathbf{u} - \nu \triangle \mathbf{u} + (\mathbf{u} \cdot \nabla)\mathbf{u} + \nabla p &= f, && \text{in} \quad \Omega \times (0,T), \\ \operatorname{div} \mathbf{u} &= 0 && \text{in} \quad \Omega \times (0,T), \\ \mathbf{u} &= 0 && \text{on} \quad \partial\Omega \times [0,T], \\ \mathbf{u}(x,0) &= \mathbf{u}_0(x) && \text{in} \quad \Omega \quad \text{for} \quad t = 0, \end{aligned} \tag{1}$$

where ν denotes the viscosity, f is the body force and $\mathbf{u}_0$ the initial velocity field at time $t = 0$. For simplicity, we restrict to 2D and we assume homogeneous Dirichlet conditions at the boundary $\partial\Omega$ of a polygonal domain Ω. To make this problem well-posed in the case of pure Dirichlet boundary conditions, we have to look for p in the subspace $L_0^2(\Omega) \subset L^2(\Omega)$ of functions with zero integral mean value. For the time discretization, we decompose the time interval $I = [0,T]$ into N subintervals

$I_n := (t_{n-1}, t_n]$, where $n = 1, \ldots, N$ and $0 = t_0 < t_1 < \cdots < t_{N-1} < t_N = T$. The symbol τ denotes the *time discretization parameter* and is also used as the maximum time step size $\tau := \max_{1 \le n \le N} \tau_n$, where $\tau_n := t_n - t_{n-1}$. Then, for the subsequent continuous and discontinuous Galerkin time stepping schemes, we approximate the solution $\mathbf{u}$ by means of a function $\mathbf{u}_\tau$ which is piecewise polynomial of order k with respect to time, i.e., we are looking for $\mathbf{u}_\tau$ in the discrete time space (with $\mathbf{V} = (H_0^1(\Omega))^2$)

$$\mathbf{X}_\tau^k := \{\mathbf{u} \in C(I, \mathbf{V}) : \mathbf{u}\big|_{I_n} \in \mathbb{P}_k(I_n, \mathbf{V}) \quad \forall\, n = 1, \ldots, N\}, \tag{2}$$

where

$$\mathbb{P}_k(I_n, \mathbf{V}) := \Big\{\mathbf{u} : I_n \to \mathbf{V} : \mathbf{u}(t) = \sum_{j=0}^{k} \mathbf{U}^j t^j \ , \ \forall\, t \in I_n, \ \mathbf{U}^j \in \mathbf{V}, \ \forall\, j\Big\}.$$

Moreover, we introduce the discrete time test space

$$\mathbf{Y}_\tau^{k-1} := \{\mathbf{v} \in L^2(I, \mathbf{V}) : \mathbf{v}\big|_{I_n} \in \mathbb{P}_{k-1}(I_n, \mathbf{V}) \quad \forall\, n = 1, \ldots, N\} \tag{3}$$

consisting of piecewise polynomials of order $k-1$ which are (globally) discontinuous at the end points of the time intervals. Similarly, we will use for the time-discrete pressure p_τ an analogous ansatz space $\tilde{X}_\tau^k$, where the vector valued space $\mathbf{V}$ is replaced by the scalar valued space $Q = L_0^2(\Omega)$, and an analogous discontinuous test space $\tilde{Y}_\tau^{k-1}$.

Now, in order to derive the time discretization, we multiply the equations in (1) with some suitable I_n-supported test functions and integrate over $\Omega \times I_n$. To determine $\mathbf{u}_\tau|_{I_n}$ and $p_\tau|_{I_n}$ we represent them by the polynomial ansatz

$$\mathbf{u}_\tau|_{I_n}(t) := \sum_{j=0}^{k} \mathbf{U}_n^j \phi_{n,j}(t), \qquad p_\tau|_{I_n}(t) := \sum_{j=0}^{k} P_n^j \phi_{n,j}(t), \tag{4}$$

where the "coefficients" $(\mathbf{U}_n^j, P_n^j)$ are elements of the function spaces $\mathbf{V} \times Q$ and the polynomial functions $\phi_{n,j} \in \mathbb{P}_k(I_n)$ are the Lagrange basis functions with respect to the $k+1$ nodal points $t_{n,j} \in I_n$ satisfying the conditions

$$\phi_{n,j}(t_{n,i}) = \delta_{i,j}, \qquad i, j = 0, \ldots, k \tag{5}$$

with the Kronecker symbol $\delta_{i,j}$. For an easy treatment of the initial condition, we set $t_{n,0} = t_{n-1}$. Then, the initial condition is equivalent to the condition

$$\mathbf{U}_n^0 = \mathbf{u}_\tau|_{I_{n-1}}(t_{n-1}) \quad \text{if} \quad n \ge 2 \qquad \text{or} \qquad \mathbf{U}_n^0 = \mathbf{u}_0 \quad \text{if} \quad n = 1. \tag{6}$$

The other points $t_{n,1},\ldots,t_{n,k}$ are chosen as the quadrature points of the k-point Gaussian formula on I_n which is exact if the function to be integrated is a polynomial of degree less or equal to $2k-1$. We define the basis functions $\phi_{n,j} \in \mathbb{P}_k(I_n)$ of (4) via affine reference transformations (see [3,5] for more details). Now, we can describe the *time discrete I_n-problem of the cGP(k)-method* [3,6]:

Find on the interval $I_n = (t_{n-1}, t_n]$ the k unknown pairs of "coefficients" $(\mathbf{U}_n^j, P_n^j) \in \mathbf{V} \times Q$, $j = 1,\ldots,k$, such that for all $i = 1,\ldots,k$, it holds for all $\mathbf{v} \in \mathbf{V}$, $q \in Q$

$$\sum_{j=0}^{k} \alpha_{i,j} \left(\mathbf{U}_n^j, \mathbf{v}\right)_{\Omega} + \frac{\tau_n}{2} a(\mathbf{U}_n^i, \mathbf{v}) + \frac{\tau_n}{2} n(\mathbf{U}_n^i, \mathbf{U}_n^i, \mathbf{v}) + \frac{\tau_n}{2} b(\mathbf{v}, P_n^i) = \tfrac{\tau_n}{2} \left(f(t_{n,i}), \mathbf{v}\right)_{\Omega}, \qquad b(\mathbf{U}_n^i, q) = 0, \tag{7}$$

with $\mathbf{U}_n^0 := \mathbf{u}_\tau(t_{n-1})$ for $n > 1$ and $\mathbf{U}_1^0 := \mathbf{u}_0$. Here, $\alpha_{i,j}$ denote some constants independent of τ_n and $(\cdot,\cdot)_\Omega$ the usual inner product in $L^2(\Omega)$. The bilinear form $a(\cdot,\cdot)$ on $\mathbf{V} \times \mathbf{V}$, $b(\cdot,\cdot)$ on $\mathbf{V} \times Q$ and the trilinear form $n(\cdot,\cdot,\cdot)$ on $\mathbf{V} \times \mathbf{V} \times \mathbf{V}$, respectively, are defined by

$$a(\mathbf{u},\mathbf{v}) := \int_\Omega \nabla \mathbf{u} \cdot \nabla \mathbf{v}\, dx, \quad b(\mathbf{v},p) := -\int_\Omega \nabla \cdot \mathbf{v}\, p\, dx, \quad n(\mathbf{w},\mathbf{u},\mathbf{v}) := \int_\Omega (\mathbf{w}\cdot\nabla)\mathbf{u} \cdot \mathbf{v}\, dx.$$

In the following, we specify the cGP(k)-method for the case $k = 2$ where the discretization error in the L^2-norm is superconvergent of order 4 at the discrete time points.

2.1 The cGP(2)-Method

Here, we use the 2-point Gaussian formula with the quadrature points $\hat{t}_1 = -\frac{1}{\sqrt{3}}$ and $\hat{t}_2 = \frac{1}{\sqrt{3}}$ at the reference interval $(-1, 1]$. Then, the coefficients in (7) are

$$(\alpha_{i,j}) = \begin{pmatrix} -\sqrt{3} & \frac{3}{2} & \frac{2\sqrt{3}-3}{2} \\ \sqrt{3} & \frac{-2\sqrt{3}-3}{2} & \frac{3}{2} \end{pmatrix} \qquad i = 1,2,\ j = 0,1,2.$$

Consequently, on the time interval I_n, we have to solve for the two "unknowns"

$$(\mathbf{U}_n^j, P_n^j) = \left(\mathbf{u}_\tau(t_{n,j}), p_\tau(t_{n,j})\right) \in \mathbf{V} \times Q \qquad \text{with} \qquad t_{n,j} := T_n(\hat{t}_j) \quad \text{for} \quad j = 1,2,$$

where $T_n : (-1,1] \to I_n$ denotes the affine reference mapping. The corresponding coupled system reads

$$\begin{aligned}
\alpha_{1,1}\left(\mathbf{U}_n^1,\mathbf{v}\right)_\Omega + \tfrac{\tau_n}{2}a(\mathbf{U}_n^1,\mathbf{v}) + \tfrac{\tau_n}{2}n(\mathbf{U}_n^1,\mathbf{U}_n^1,\mathbf{v}) + \alpha_{1,2}\left(\mathbf{U}_n^2,\mathbf{v}\right)_\Omega + \tfrac{\tau_n}{2}b(\mathbf{v},P_n^1) &= \ell_1(\mathbf{v})\\
\alpha_{2,1}\left(\mathbf{U}_n^1,\mathbf{v}\right)_\Omega + \alpha_{2,2}\left(\mathbf{U}_n^2,\mathbf{v}\right)_\Omega + \tfrac{\tau_n}{2}a(\mathbf{U}_n^2,\mathbf{v}) + \tfrac{\tau_n}{2}n(\mathbf{U}_n^2,\mathbf{U}_n^2,\mathbf{v}) + \tfrac{\tau_n}{2}b(\mathbf{v},P_n^2) &= \ell_2(\mathbf{v})\\
b(\mathbf{U}_n^1,q) &= 0\\
b(\mathbf{U}_n^2,q) &= 0,
\end{aligned} \tag{8}$$

which has to be satisfied for all $\mathbf{v}\in\mathbf{V}$ and $q\in Q$ with the linear functionals

$$\ell_i(\mathbf{v}) := \frac{\tau_n}{2}\left(f(t_{n,i}),\mathbf{v}\right)_\Omega - \alpha_{i,0}\left(\mathbf{U}_n^0,\mathbf{v}\right)_\Omega \qquad i=1,2. \tag{9}$$

Once we have determined the solutions $(\mathbf{U}_n^j, P_n^j)$ at the Gaussian points in the interior of the interval I_n, we get the solution at the right boundary t_n of I_n by means of quadratic extrapolation from the ansatz (4), i.e.,

$$\mathbf{u}_\tau(t_n) = \mathbf{U}_n^0 + \sqrt{3}(\mathbf{U}_n^2 - \mathbf{U}_n^1), \tag{10}$$

where $\mathbf{U}_n^0$ is the initial value at the time interval I_n.

After discretizing Eq. (1) in time, we now discretize the resulting "I_n-problems" (8) in space by using the finite element method [2,7] with the well-known Q_2/P_1^{disc} element. Here, we present only the resulting block system for the nodal vectors $\underline{\mathbf{U}}_n^j = (\underline{U}_n^j, \underline{V}_n^j)$ and $\underline{P}_n^j$, $j=1,2$, associated with the finite element functions that approximate the functions $\mathbf{U}_n^j\in\mathbf{V}$ and $P_n^j\in Q$ in (8). The 6×6 block system on the time interval I_n reads:

For given initial velocity $\underline{\mathbf{U}}_n^0 = (\underline{U}_n^0, \underline{V}_n^0)$, find the nodal vectors $\underline{\mathbf{U}}_n^j = (\underline{U}_n^j, \underline{V}_n^j)$ and $\underline{P}_n^j$, $j=1,2$, such that for

$$u = (\underline{U}_n^1, \underline{U}_n^2), \quad v = (\underline{V}_n^1, \underline{V}_n^2), \quad p = (\tau_n\underline{P}_n^1, \tau_n\underline{P}_n^2),$$

it holds

$$\begin{bmatrix} A(u,v) & 0 & B_u\\ 0 & A(u,v) & B_v\\ B_u^T & B_v^T & 0\end{bmatrix}\begin{bmatrix} u\\ v\\ p\end{bmatrix} = \begin{bmatrix} R_u\\ R_v\\ 0\end{bmatrix} \tag{11}$$

where

$$A(u,v) = \begin{bmatrix} 3M+\tau_nL+\tau_nN_1(u,v) & \left(2\sqrt{3}-3\right)M\\ \left(-2\sqrt{3}-3\right)M & 3M+\tau_nL+\tau_nN_2(u,v)\end{bmatrix},$$

$$B_u = \begin{bmatrix} B_1 & 0\\ 0 & B_1\end{bmatrix}, \qquad B_v = \begin{bmatrix} B_2 & 0\\ 0 & B_2\end{bmatrix}$$

and $N_1(u,v) = N(\underline{U}_n^1, \underline{V}_n^1)$, $N_2(u,v) = N(\underline{U}_n^2, \underline{V}_n^2)$ correspond to the nonlinear operators evaluated at the Gauß points $t_{n,1}$ and $t_{n,2}$, respectively.

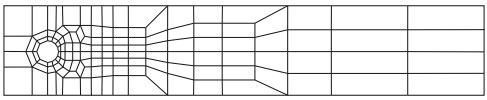


Fig. 1 Coarse mesh for *flow around cylinder*

Fig. 2 Size of the different systems in space

Lev.	#EL	#DOF(total)
2	520	5 928
3	2 080	23 296
4	8 320	92 352

Here, M, L, B_1, B_2 denote the mass, Laplacian and pressure matrices, respectively. The right hand side vectors $R_u = (R_u^1, R_u^2)$ and $R_v = (R_v^1, R_v^2)$ are given by

$$R_u^j := \tau_n F_n^j - 2\sqrt{3}(-1)^j M \underline{U}_n^0, \qquad R_v^j := \tau_n G_n^j - 2\sqrt{3}(-1)^j M \underline{V}_n^0, \qquad j = 1, 2,$$

where F_n^j and G_n^j are the vectors corresponding to the term $\big(f(t_{n,j}), \mathbf{v}\big)_\Omega$ with test functions $\mathbf{v} = (\varphi_i, 0)$ and $\mathbf{v} = (0, \varphi_i)$, respectively ($\varphi_i$ denoting the scalar basis functions for velocity).

Once we have determined the solutions $\underline{\mathbf{U}}_n^j = (\underline{U}_n^j, \underline{V}_n^j)$, $j = 1, 2$, we compute the nodal vector $\underline{\mathbf{U}}_{n+1}^0 = (\underline{U}_{n+1}^0, \underline{V}_{n+1}^0)$ of the fully discrete solution $\mathbf{u}_{\tau,h}(t_n) = (u_{\tau,h}(t_n), v_{\tau,h}(t_n))$ at time t_n by using the following quadratic extrapolation

$$u_{\tau,h}(t_n) \sim \underline{U}_{n+1}^0 := \underline{U}_n^0 + \sqrt{3}(\underline{U}_n^2 - \underline{U}_n^1), \quad v_{\tau,h}(t_n) \sim \underline{V}_{n+1}^0 := \underline{V}_n^0 + \sqrt{3}(\underline{V}_n^2 - \underline{V}_n^1).$$

3 Numerical Results

In this section, we perform nonstationary simulations to demonstrate the temporal accuracy and efficiency of the presented time discretization scheme. As a test problem, we consider the *flow around cylinder* which exactly corresponds to the benchmark configuration in [8]. In this simulation, we concentrate only on the nonstationary behavior of the flow pattern with periodic oscillations and examine the ability of the presented time discretization scheme to capture the dynamics of the flow. Details regarding the benchmark settings can be found at www.featflow.de/en/benchmarks/cfdbenchmarking.html. The examined accuracy of the benchmark crucially depends on the following quantities

$$F_D = \int_S (\rho\nu \frac{\partial u_t}{\partial n} n_y - p n_x) dS, \quad \text{and particularly} \quad F_L = -\int_S (\rho\nu \frac{\partial u_t}{\partial n} n_x + p n_y) dS$$

representing the total forces in the horizontal and vertical directions, respectively.

Figure 1 shows the initial coarse mesh (level 1), which will be uniformly refined, and Fig. 2 presents for different space mesh levels the number '#EL' of elements

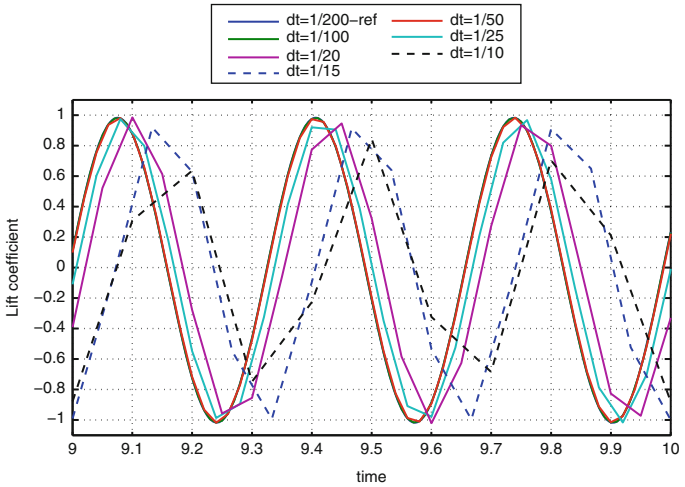


Fig. 3 Lift coefficient for different Δt using cGP(2)-method at space mesh level 4

Table 1 *Deviation per cycle (in %) of lift values* after appr. 30 cycles at different space level

τ	Lev=2	Lev=3	Lev=4
1/100	0.00	0.00	0.00
1/50	0.01	0.01	0.01
1/25	0.11	0.13	0.12
1/20	0.25	0.29	0.29
1/15	0.72	0.84	0.85
1/10	0.93	0.96	0.98

and the total number '#DOF' of degrees of freedom. In order to demonstrate the accuracy of the higher order time discretization, the flow is started from the same developed solution at time t_0, and the simulation is performed until $T = 10$ for various uniform time step sizes $\tau_n := \tau$. After $T = 10$, all the introduced quantities are plotted and analyzed in detail. Here, we will concentrate on the values of the lift coefficient (C_L). To this end, we show in Fig. 3 only the zoomed picture in the last time unit from $T = 9$ to $T = 10$ at space level 4. We can see that, except for the dashed lines (dt >= 1/15), all curves are pretty close to the reference curve (dt=1/200) even after 30 periods of cycles with a length of about 0.33.

Based on these results, we present a more quantitative analysis, too. For different time step sizes τ, Table 1 shows the 'deviation in percentage per cycle' from the reference values, i.e., $\frac{\Delta x}{30 \times 0.33} \times 100\,\%$, where Δx is the total deviation after $T = 10$ (with length of period ≈ 0.33, number of cycles until $T = 10 \approx 30$). We can see that the maximum time step size to gain the accuracy with an error of less than 1 % per period for the cGP(2)-method is $\tau = 1/10$.

Remark 1. It has been analyzed in [4], that the cGP(1) or Crank-Nicolson scheme gains similar results with $\tau = 1/100$.

Table 2 Averaged number of nonlinear iterations (#NL) per time step and multigrid linear iterations (#MG) per nonlinear step

	FP	NWT	FP	NWT	FP	NWT	FP	NWT	FP	NWT	FP	NWT
	#NL	#NL	#NL	#NL	#NL	#NL	#MG	#MG	#MG	#MG	#MG	#MG
τ	Level = 3		Level = 4		Level = 5		Level = 3		Level = 4		Level = 5	
1/10	11.18	5.00	10.18	4.91	9.00	4.55	10.91	10.60	10.20	9.60	9.11	9.00
1/15	8.62	4.06	7.94	4.00	7.00	4.00	10.89	10.20	10.50	10.00	9.43	9.25
1/20	7.14	4.00	7.00	4.00	6.00	3.90	11.00	10.75	10.71	10.25	9.67	9.25
1/25	7.00	4.00	6.00	3.04	5.04	3.04	11.43	11.00	11.00	10.25	9.50	9.25
1/50	5.02	3.02	4.24	3.00	4.02	3.00	11.83	11.00	11.60	10.67	10.00	9.67
1/100	4.01	2.01	3.98	2.01	3.01	2.01	11.60	11.00	12.00	11.33	11.00	10.33
1/200	3.00	2.00	3.00	2.00	3.00	2.00	12.00	11.00	12.50	12.00	12.00	11.33

4 Solver Analysis

In order to measure the efficiency of the nonlinear solver for the presented time discretization scheme, we show in Table 2 the averaged number '#NL' of nonlinear iterations per time step for the fixed-point (FP) and Newton (NWT) method on different space mesh levels. To analyze the corresponding behavior of the multigrid solver for the solution of linear subproblems, we present the averaged number '#MG' of multigrid iterations per (nonlinear) step. Here, the multigrid solver uses a preconditioned GMRES method (preconditioned with a cell oriented Vanka scheme) as smoother and applies four pre- and post-smoothing steps. Table 2 reveals that, for both nonlinear solver methods, almost the same number of iterations are required if τ is fixed and the space mesh level increases. Moreover, for fixed space mesh level, the number of nonlinear iterations decreases if the time step size is reduced, as expected. Concerning the number of nonlinear iterations, Table 2 shows that the Newton method is more efficient than the fixed point iteration. We also see that the number of multigrid iterations remains fairly constant if we increase the refinement level of the space mesh. There is also no noticeable increase in the number of iterations if we decrease the time step size. This means that the behavior of the multigrid solver is almost independent of the space mesh size and the time step size.

5 Conclusion

We have implemented the cGP(2)-method for the nonstationary Navier-Stokes equations. The spatial discretization is carried out by using biquadratic finite elements for velocity and discontinuous linear elements for pressure. The discretized systems of nonlinear equations are treated by means of the fixed-point and the Newton method. The associated linear systems have been solved using a geometrical multigrid method with the Vanka-type preconditioned GMRES method as smoother. From the numerical studies, we observe that the cGP(2)-scheme provides highly accurate

numerical solutions at quite large time step sizes. Moreover, the analysis of the numerical costs shows that the arising nonlinear block-systems in the implicit time discretization scheme can be solved very efficiently with nearly optimal complexity.

Acknowledgements The authors want to express their gratitude to the German Research Association (DFG) and the Higher Education Commission (HEC) of Pakistan for their financial support of the study; contract/grant number: SCHI 576/2-1, TU 102/35-1 and LC06052 by MSMT.

References

1. Aziz, A.K., Monk, P.: Continuous finite elements in space and time for the heat equation. Math. Comp. **52**(186), 255–274 (1989)
2. C. Cuvelier, A. Segal, A. A. Steenhoven: Finite element methods and Navier-Stokes equations. D. Reidel Publishing Company (1986)
3. Hussain, S., Schieweck, F., Turek, S.: Higher order Galerkin time discretizations and fast multigrid solvers for the heat equation. Journal of Numerical Mathematics **19**(1), 41–61 (2011)
4. Hussain, S., Schieweck, F., Turek, S.: A new efficient and stable finite element solver of higher order in space and time for nonstationary incompressible flows. International Journal for Numerical Methods in Fluids (2012). Submitted for publication
5. Hussain, S., Schieweck, F., Turek, S.: A note on accurate and efficient higher order Galerkin time stepping schemes for the nonstationary Stokes equations. The Open Numerical Methods Journal **4**, 35–45 (2012)
6. Schieweck, F.: A-stable discontinuous Galerkin-Petrov time discretization of higher order. J. Numer. Math. **18**(1), 25–57 (2010)
7. Thomée, V.: Galerkin finite element methods for parabolic problems, *Springer Series in Computational Mathematics*, vol. 25, second edn. Springer-Verlag, Berlin (2006)
8. Turek, S., Schäfer, M.: Benchmark computations of laminar flow around cylinder. In: E. Hirschel (ed.) Flow Simulation with High-Performance Computers II, *Notes on Numerical Fluid Mechanics*, vol. 52, pp. 547–566. Vieweg (1996)

On the Density-Enthalpy Method for the $2D$ Darcy Flow

D. Ibrahim, F.J. Vermolen, and C. Vuik

Abstract The new approach, called the density-enthalpy method, has certain advantages over the tradition methods for solving multi-phase flow problems. The system is modeled by the mass and energy conservation, Darcy's law, and other thermodynamic relations. It is solved by using the standard Galerkin algorithm for spatial discretization, backward Euler for time integration, and Newton-Raphson iteration for linearization. In this paper, the main objective is to study the effect of gravity on Darcy flow.

1 Introduction

Multi-phase flow occurs in many natural and industrial processes [9]. Underground flow of hydrocarbons, boilers, food processing and plastic industry are few examples which involve multi-phase flow as well as porous media. Traditional methods to solve such systems include moving-boundary, level-set and phase-field methods. An alternative (and relatively new) approach is to use pre-computed phase diagrams to define system variables (e.g., temperature T $[K]$, pressure P $[P_a]$, gas mass-fraction X_G $[K_g/K_g]$, etc) in terms of two state-variables, density ρ $[k_g/m^3]$ and specific enthalpy h $[J/k_g]$. We call it the density-enthalpy method (or simply ρ-h method). The method is first presented by R. Arendsen in [1] for a spatially homogeneous system. In this approach, the flow system is modeled by a set of PDEs (the mass and energy conservation and Darcy's law) and other relations. The system of equations is solved for ρ and h (henceforth called enthalpy). Other solution variables (like T, P, X_G, etc) are obtained by using the available phase diagrams.

D. Ibrahim (✉) · F.J. Vermolen · C. Vuik
Delft Institute of Applied Mathematics, Delft University of Technology, Mekelweg 4, 2628 CD, Delft, The Netherlands
e-mail: D.Ibrahim@tudelft.nl; F.J.Vermolen@tudelft.nl; C.Vuik@tudelft.nl

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_55,

The density-enthalpy method offers several advantages over traditional methods. The solution algorithm does not require an explicit phase-front tracking. Other methods need phase-tracking because they use separate sets of equations for gas, liquid and solid phases. In contrast, ρ-h approach uses the same set of equations for all phases (for any one or a mixer of the phases). The lack of phase-front tracking requirement saves a significant amount of computation time. Hence, it is potentially an efficient (faster) method. Moreover, it is purely a physical approach (uses thermodynamics) and does not use abstract mathematical parameters.

We solve the system numerically by using the standard Galerkin algorithm for the spatial discretization. For time quadrature, we use backward Euler scheme together with the Newton-Raphson iteration. For other linearization schemes, we refer to [2,3] and [5,6]. The primary objective of this paper is to show that the ρ-h approach works well for a two-phase flow due to gravity (alongside the other driving factor i.e. the temperature gradient). We use a $2D$ ρ-h model as given in [6]. However, we add a gravity term in the Darcy's law.

2 The Density-Enthalpy Model

We consider the flow of Propane (C_3H_8) in a unit square Ω. The system is modeled by the following nonlinearly coupled PDEs and other thermodynamics relations

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \left(\rho \boldsymbol{v}\right) = 0, \quad \boldsymbol{x} \in \Omega, \quad t > 0 \ \text{(mass conservation)}, \tag{1}$$

$$\frac{\partial s}{\partial t} + \nabla \cdot (s\boldsymbol{v}) - \nabla \cdot (D\nabla T) = q, \quad \boldsymbol{x} \in \Omega, \quad t > 0 \ \text{(energy conservation)}, \tag{2}$$

$$\boldsymbol{v} + \frac{K}{\mu}\nabla P + \frac{K}{\mu} g\rho \hat{\boldsymbol{y}} = 0, \quad \boldsymbol{x} \in \Omega, \quad t > 0 \ \text{(Darcy's law)}, \tag{3}$$

$$T = T(\rho, h), \quad \boldsymbol{x} \in \Omega, \quad t > 0 \ \text{(thermodynamical relation)}, \tag{4}$$

$$P = P(\rho, h), \quad \boldsymbol{x} \in \Omega, \quad t > 0 \ \text{(thermodynamical relation)}, \tag{5}$$

$$s = \rho h, \quad \boldsymbol{x} \in \Omega, \quad t > 0 \ \text{(total enthalpy)}, \tag{6}$$

$$X_G = X_G(\rho, h), \quad \boldsymbol{x} \in \Omega, \quad t > 0 \ \text{(thermodynamical relation)}. \tag{7}$$

The permeability K, dynamic viscosity μ, and heat diffusivity D are constants because only one substance is present in a homogeneous porous medium. The velocity $\boldsymbol{v}$ and the operator ∇ are also $2D$ with components $< v_x, v_y >$ and $< \frac{\partial}{\partial x}, \frac{\partial}{\partial y} >$, respectively. We take the gravity along the y-axis, indicated by a unit vector $\hat{\boldsymbol{y}}$ in that direction. The initial and boundary conditions are given as follows

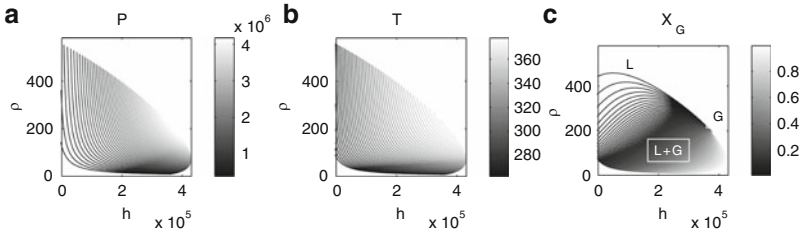


Fig. 1 Numerical ρ-h diagrams. Plots of (**a**) isotherms, (**b**) isobars, and (**c**) constant gas mass fraction curves

$$T(\boldsymbol{x},0) = T_0(\boldsymbol{x}), \qquad \boldsymbol{x} \in \Omega,$$

$$X_G(\boldsymbol{x},0) = X_{G,0}(\boldsymbol{x}), \qquad \boldsymbol{x} \in \Omega,$$

$$\rho \boldsymbol{v} \cdot \hat{\boldsymbol{n}} = 0, \qquad \boldsymbol{x} \in \Gamma, \quad t > 0 \quad \text{(zero mass flux)}, \tag{8}$$

$$-D\frac{\partial T}{\partial \hat{\boldsymbol{n}}} + s\boldsymbol{v} \cdot \hat{\boldsymbol{n}} = 0, \qquad \boldsymbol{x} \in \Gamma, \quad t > 0 \quad \text{(zero energy flux)}, \tag{9}$$

where $\hat{\boldsymbol{n}}$ is a unit normal to the boundary. For nonzero boundary fluxes, we refer to [7].

In this paper we only deal with the two-phase fluid flow (gas + liquid) and a fixed volume isolated system. For the fluid flow, the system is unconditionally stable. The stability and error analysis of the ρ-h model is discussed in [7].

2.1 Numerical Density-Enthalpy Phase Diagrams

The pre-computed density-enthalpy phase diagrams, given in Fig. 1, provide a transformation from (ρ, h) to (T, P, X_G), therefore enabling to use ρ and h as our state variables. The computation of these diagrams is discussed in [1] and [6]. Other phase diagrams for a transformation from (T, X_G) to (ρ, h, P) are also available. This means thats we can specify initial conditions in terms of T and X_G, which is more suitable in practice.

3 The Solution Strategy

For the numerical solution of the given system, we use the standard Galerkin algorithm for spatial discretization, backward Euler for time integration and Newton-Raphson for the linearization. In other words, a fully-implicit time integration scheme is utilized. We provide some detail about the numerical solution of the mass equation (1) but the other equations are treated very briefly.

3.1 Numerical Treatment of the Mass Equation

The weak-form of Eq. (1) is obtained by multiplying it by a bilinear test function ϕ and integrating over Ω. This gives

$$\frac{1}{\Delta t}\int_{\Omega}(\rho^{\tau}-\rho^{\tau-1})\phi \, d\Omega + \int_{\Omega}\nabla\cdot(\rho^{\tau}\boldsymbol{v}^{\tau})\phi \, d\Omega = 0, \tag{10}$$

where τ is a time index. By using the vector product rule and the divergence theorem, we obtain

$$\frac{1}{\Delta t}\int_{\Omega}(\rho^{\tau}-\rho^{\tau-1})\phi d\Omega + \int_{\Gamma}\rho^{\tau}\boldsymbol{v}^{\tau}\phi\cdot\hat{\boldsymbol{n}}d\Gamma - \int_{\Omega}\rho^{\tau}\boldsymbol{v}^{\tau}\cdot\nabla\phi d\Omega = 0. \tag{11}$$

We apply the boundary conditions, which makes the second integral equal to zero, and then linearize this equation about $\rho^{\tau,k}$ and $\boldsymbol{v}^{\tau,k}$. The resulting nonlinear partial differential equation is solved by a Newton method based on successive linearization of the PDE. This gives

$$\begin{aligned}&\frac{1}{\Delta t}\int_{\Omega}\delta\rho\,\phi(\boldsymbol{x})\,d\Omega + \frac{1}{\Delta t}\int_{\Omega}(\rho^{k}-\rho^{\tau-1})\phi(\boldsymbol{x})\,d\Omega\\ &-\int_{\Omega}(\rho^{k}v_x^{k}+v_x^{k}\delta\rho+\rho^{k}\delta v_x)\frac{\partial\phi}{\partial x}\,d\Omega - \int_{\Omega}(\rho^{k}v_y^{k}+v_y^{k}\delta\rho+\rho^{k}\delta v_y)\frac{\partial\phi}{\partial y}\,d\Omega = 0,\end{aligned}$$

where $\delta\rho = \rho^{k+1}-\rho^{k}$, $\delta v_x = v_x^{k+1}-v_x^{k}$, and $\delta v_y = v_y^{k+1}-v_y^{k}$. For a better readability, we omit the index τ but it is understood to be there (e.g., ρ^{k} actually represents $\rho^{\tau,k}$). We choose a set of basis functions $\{\phi_i\}_N$ and apply the approximation $\delta\rho(\boldsymbol{x},\tau) \approx \sum_{j=1}^{N}\delta\rho_j(\tau)\,\phi_j(\boldsymbol{x})$, $\delta v_x(\boldsymbol{x},\tau) \approx \sum_{j=1}^{N}\delta v_{x,j}(\tau)\,\phi_j(\boldsymbol{x})$, and $\delta v_y(\boldsymbol{x},\tau) \approx \sum_{j=1}^{N}\delta v_{y,j}(\tau)\,\phi_j(\boldsymbol{x})$. After substitution into the weak form and choosing $\phi=\phi_i$, for $i=1,\ldots,N$, we obtain

$$\begin{aligned}&\frac{1}{\Delta t}\sum_{j=1}^{N}\delta\rho_j\int_{\Omega}\phi_i\phi_j\,d\Omega + \frac{1}{\Delta t}\int_{\Omega}(\rho^{k}-\rho^{\tau-1})\phi_i\,d\Omega\\ &-\int_{\Omega}\rho^{k}v_x^{k}\frac{\partial\phi_i}{\partial x}d\Omega - \sum_{j=1}^{N}\delta\rho_j\int_{\Omega}v_x^{k}\phi_j\frac{\partial\phi_i}{\partial x}\,d\Omega - \sum_{j=1}^{N}\delta v_{x,j}\int_{\Omega}\rho^{k}\phi_j\frac{\partial\phi_i}{\partial x}\,d\Omega\\ &-\int_{\Omega}\rho^{k}v_y^{k}\frac{\partial\phi_i}{\partial y}d\Omega - \sum_{j=1}^{N}\delta\rho_j\int_{\Omega}v_y^{k}\phi_j\frac{\partial\phi_i}{\partial y}\,d\Omega - \sum_{j=1}^{N}\delta v_{y,j}\int_{\Omega}\rho^{k}\phi_j\frac{\partial\phi_i}{\partial y}\,d\Omega = 0.\end{aligned} \tag{12}$$

These integrals are determined by using an iso-parametric transformation [3] together with the Newton-Cotes quadrature rule for a unit square.

$$\int_0^1 \int_0^1 I(\xi, \eta)\, dx\, dy = \sum_{i=1}^{4} I(\xi_i, \eta_i),$$

where (ξ_i, η_i) are the vertices of the reference element in some (ξ, η)-plane. This transformation allows us to use generic convex quadrilateral elements. For a complete numerical treatment of Eqs. (1)–(6), we refer to [4]. Let the weak-forms for the PDEs and other relations be arranged in the homogeneous form

$$\boldsymbol{F}(\boldsymbol{G}) = \boldsymbol{0},$$

where $\boldsymbol{F}$ and $\boldsymbol{G}$ are vectors of the same dimension. The vector $\boldsymbol{G}$ is the required solution set and $\boldsymbol{F}$ represents the set of relations between them. The Taylor expansion of $\boldsymbol{F}(\boldsymbol{G})$ about $\boldsymbol{G}^k$ (for some k) is expressed as

$$\boldsymbol{F}(\boldsymbol{G}^k) + \left.\frac{\partial \boldsymbol{F}}{\partial \boldsymbol{G}}\right|_{\boldsymbol{G}=\boldsymbol{G}^k} \boldsymbol{\delta G} + \frac{1}{2}\, \boldsymbol{\delta G}^T \left.\frac{\partial^2 \boldsymbol{F}}{\partial \boldsymbol{G}^2}\right|_{\boldsymbol{G}=\boldsymbol{G}^k} \boldsymbol{\delta G} + \ldots = \boldsymbol{0},$$

where $\boldsymbol{\delta G} = \boldsymbol{G}^{k+1} - \boldsymbol{G}^k$ and $J = \left.\frac{\partial \boldsymbol{F}}{\partial \boldsymbol{G}}\right|_{\boldsymbol{G}=\boldsymbol{G}^k}$ is the Jacobian. Assuming that $\boldsymbol{F}(\boldsymbol{G})$ is sufficiently smooth in the neighborhood of $\boldsymbol{G}^k$, we approximate $\boldsymbol{F}(\boldsymbol{G})$ by the first two terms of its Taylor's expansion. This gives

$$\boldsymbol{F}(\boldsymbol{G}^k) + J\,(\boldsymbol{G}^{k+1} - \boldsymbol{G}^k) = \boldsymbol{0}, \qquad \text{or} \qquad \boldsymbol{G}^{k+1} = \boldsymbol{G}^k - J^{-1}\boldsymbol{F}(\boldsymbol{G}^k).$$

Hence, the linearization actually leads to a Newton-Raphson iteration, making k as index of the Newton loop.

4 Numerical Experiments

The parameters used in this section are provided in Table 1, where N_x and N_y are the number of nodes in the directions indicated by the subscripts. The values of μ, K, and D are taken from [6]. The simulations results are interpreted on the basis of physical laws. For example, the flow of heat energy from a point at higher temperature to a point at lower temperature is consistent with the Fourier law of heat conduction. Similarly, the number of state variables, in our case, are governed by Gibbs phase rule.

Subject to the conditions of thermal equilibrium, the Gibbs phase rule [8] is given by the relation $F = C - \Phi + 2$, where F is the number of degrees of freedom C

Table 1 Parameters used in the simulations

N_x	N_y	Δt $[s]$	μ $[P_a.s]$	K $[m^2]$	D $[W/m/K]$	g $[m/s^2]$
100	100	1/10	5×10^{-5}	10^{-10}	0.05	9.81

is the number of components (or substances), and Φ is the number of phases in thermodynamic equilibrium with each other.

In our case, $C = 1$ (one substance i.e., Propane), $\Phi = 2$ and hence, $F = 1$. This implies that one state variable is sufficient to completely describe the system. In general, we have two state variables, ρ $[k_g/m^3]$ and h $[J/k_g]$. For an isolated system, the flow in Ω definitely alters the density but the enthalpy (i.e., mass to energy ratio) remains constant. This is possible only if the heat energy (or temperature) vary accordingly. We conclude that for the isolated system, an increase in the density at any point will always be accompanied with an increase in the temperature at that point.

4.1 Flow Due to the Gravity Term

Here, we set the initial conditions such that the flow is triggered due to the gravity. These are given by

$$X_G(\boldsymbol{x}, 0) = 0.1, \qquad \text{and} \qquad T(\boldsymbol{x}, 0) = 300.$$

Since $\Delta T = 0$, this makes $\Delta P = 0$ (refer to [4]). Therefore, the velocity is initiated by the gravity term (see (3)). The plots of the solution variables are given in Fig. 2. Note that Fig. 2a depicts graphs of T for the entire spatial domain at $t = 0$ and $1.0\,[s]$. These plots indicate that the flow is only in the y-direction. The plots for v_y (at $t = 0$, 0.25, and $1.5\,[s]$) are given in Fig. 2b. From Fig. 2c–f, the cross-sections of the solution variables (along y-axis) are plotted for the entire process time. We do not have a flow in the x-direction. Hence, $v_x(\boldsymbol{x}, t) = 0$, $\boldsymbol{x} \in \Omega$, $t > 0$. The velocity along y-axis starts with $v_y(\boldsymbol{x}, 0) = -g\rho$ but $\boldsymbol{v} \cdot \hat{\boldsymbol{n}}$ immediately goes to zero because the system is isolated. This is shown in Fig. 2b, d. A flow in the y-direction causes an increase in the density, near $y = 0$ boundary (Fig. 2e) and an increase in the gas mass fraction near $y = 1$ boundary (Fig. 2f). For the temperature evolution (Fig. 2c), the Gibbs phase rule dictates that an increase in the density at a point should accompany an increase in temperature at that point. An increase in $\frac{\partial T}{\partial y}$ increases $\frac{K}{\mu}\frac{\partial P}{\partial y}$ [4]. After a certain time period, $-\frac{K}{\mu}\frac{\partial T}{\partial y}$ cancels out the term, $-g\rho$, in the Darcy's law and hence results in a zero v_y (actually this happens asymptotically). This is also a steady-state of the system. We conclude that the model and the solution algorithm works fine when the gravity term is added in the Darcy's law.

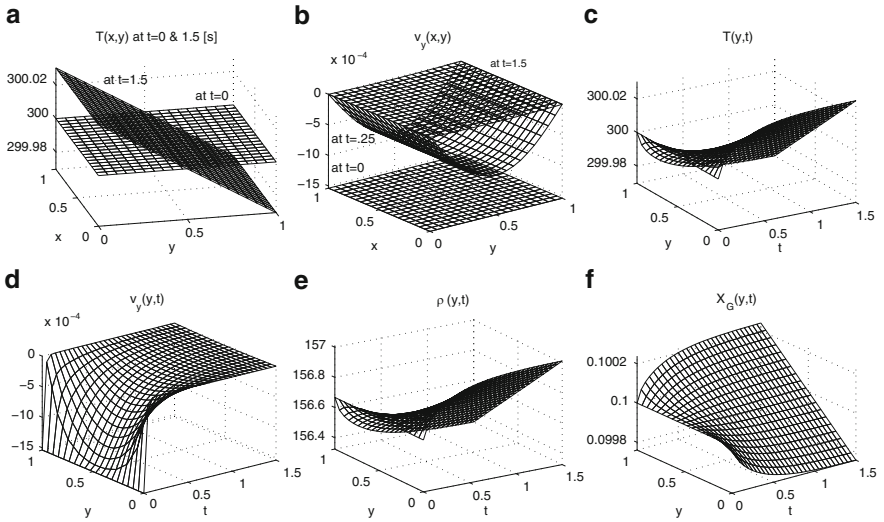


Fig. 2 Flow due to gravity term. Plots of system variables

4.2 Flow Due to the Gravity and Temperature Gradient

In this example, we show that gravity term contributes very little at the initial stages, as compare to a velocity contribution due to ΔT of few degrees $[K/m]$. We set the following initial conditions given by

$$T(\boldsymbol{x},0) = \frac{1}{2}(T_1 + T_2),$$

$$T_1 = \begin{cases} 305 & \text{for } x \in [0, 0.05],\ y \in [0,1], \\ 305 - \frac{200}{9}x + \frac{20}{18} & \text{for } x \in]0.05, 0.95],\ y \in [0,1], \\ 285 & \text{for } x \in]0.95, 1],\ y \in [0,1], \end{cases}$$

$$T_2 = \begin{cases} 305 & \text{for } y \in [0, 0.05],\ x \in [0,1], \\ 305 - \frac{200}{9}y + \frac{20}{18} & \text{for } y \in]0.05, 0.95],\ x \in [0,1], \\ 285 & \text{for } y \in]0.95, 1],\ x \in [0,1]. \end{cases}$$

$$X_G(x,0) = \begin{cases} G_1, & \{(x,y)|r < r_1\}, \\ mr - mr_1 + G_1, & \{(x,y)|r_1 \le r \le r_2\}, \\ G_2 & \{(x,y)|r < r_1\}, \end{cases} \tag{13}$$

Table 2 Parameters used in the initial conditions

r_1	r_2	r	G_1	G_2	m
0.4	0.6	$\sqrt{x^2+y^2}$	0.1	0.2	$\frac{G_1-G_2}{r_2-r_1}$

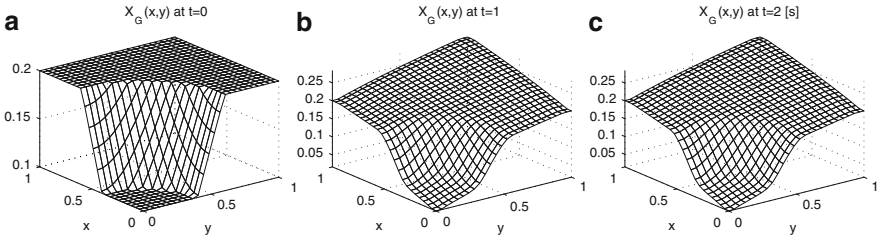


Fig. 3 Flow due to gravity and temperature gradient. Plots of X_G

where the parameters, G_i, r_i, and m are given in Table 2. Such $T(\boldsymbol{x}, 0)$ insures $\boldsymbol{v}(\boldsymbol{x}, 0) \cdot \hat{\boldsymbol{n}} = 0$ at Γ and an initial velocity in the $x = y$ direction with Ω. We give X_G plots in Fig. 3 for at three time instances. By an eyeball norm, the results are symmetric and hence, the gravity (acting in the y-direction) did not contribute much.

5 Conclusions

We conclude that the density-enthalpy model can be used to solve a multi-phase flow system where the gravity is also taken into account. In general, the gravity term in the Darcy's law plays a very little role, where we have some temperature gradient, but it is more prominent in the absence of an initial temperature gradient.

References

1. A. R. J. Arendsen, A. I. van Berkel, A. B. M. Heesink, and G. F. Versteeg. Dynamic modelling of thermal processes with phase transitions by means of a density-enthalpy phase diagram. 7th World Congress of Chemical Engineering, Glasgow, Scotland, 2005.
2. Ibrahim, F. J. Vermolen, and C. Vuik. Application of the numerical density-enthalpy method to the multi-phase flow through a porous medium. Procedia Computer S, cience 1 (2010) 781–790, Amsterdam, The Netherlands, 2010.
3. Ibrahim, C. Vuik, F.J. Vermolen, D. Hegen. Numerical Methods for Industrial Flow Problems. Delft University of Technology, Report 09–10, 2009.
4. Ibrahim, C. Vuik, F.J. Vermolen. Numerical Methods for Industrial Flow Problems. Delft University of Technology, Report 10–23, 2010.
5. Ibrahim, C. Vuik, F.J. Vermolen. Numerical Methods for Industrial Flow Problems. Delft University of Technology, Report 11–03, 2011.

6. A. Abouhafç. Finite Element Modeling Of Thermal Processes With Phase Transitions. Master Thesis, Delft University of Technology, 2007.
7. Ibrahim. Density-enthalpy method for multiphase flow problems. Dissertation, Delft University of Technology, 2012.
8. A. Faghri, Y. Zhang. Transport phenomena in multiphase systems. Elsevier Inc., pp. 111, 2006.
9. J. Bear. Dynamics of fluids in porous media. American Elsevier Publishing Company Inc., 1972.

Numerical Study of Effect of Stress Tensor for Viscous and Viscoelastic Fluids Flow

R. Keslerová

Abstract This work deals with the numerical simulation of viscous and viscoelastic fluids flow. The governing system of equations is based on the system of balance laws for mass and momentum for incompressible laminar fluids. Two models for the stress tensor are tested. For viscous fluids flow Newtonian model is used. By the combination of Newtonian and simple viscoelastic (Maxwell) models the behaviour of the mixture of viscous and viscoelastic fluids can be described. This model is called Oldroyd-B model. Both presented models (Newtonian and Oldroyd-B) can be generalized for the numerical modelling of the generalized Newtonian and Oldroyd-B fluids flow. In this case the viscosity is no more constant but is defined as a shear rate dependent viscosity function $\mu(\dot{\gamma})$. One of the most frequently used shear-thinning models is the generalized cross model. Numerical solution of the described models is based on cell-centered finite volume method using explicit Runge–Kutta time integration. Steady state solution is achieved for $t \to \infty$. In this case the artificial compressibility method can be applied. The numerical results of generalized Newtonian and generalized Oldroyd-B fluids flow obtained by this method are presented and compared.

1 Mathematical Model

The governing system of equations is the system of balance laws of mass and momentum for incompressible fluids, [2, 5, 10]:

$$\operatorname{div} \boldsymbol{u} = 0 \tag{1}$$

R. Keslerová (✉)
Faculty of Mechanical Engineering, Department of Technical Mathematics, Czech Technical University, Karlovo nám. 13, 121 35, Praha 2, Czech Republic
e-mail: keslerov@marian.fsik.cvut.cz

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_56,

$$\rho \frac{\partial \boldsymbol{u}}{\partial t} + \rho(\boldsymbol{u}.\nabla)\boldsymbol{u} = -\nabla P + \operatorname{div} \mathbf{T} \tag{2}$$

where P is the pressure, ρ is the constant density, $\boldsymbol{u}$ is the velocity vector, $\boldsymbol{u} = (u, v, w)^T$. The symbol $\mathbf{T}$ represents the stress tensor,

$$\mathbf{T} = \begin{pmatrix} t_1 & t_2 & t_3 \\ t_2 & t_4 & t_5 \\ t_3 & t_5 & t_6 \end{pmatrix} \tag{3}$$

1.1 Stress Tensor

In this work the different choices of the definition of the stress tensor are used.

(a) **Viscous fluids**
The commonly used model corresponding to Newtonian fluid is *Newtonian model*:

$$\mathbf{T} = 2\mu \mathbf{D} \tag{4}$$

where μ is dynamic viscosity and tensor $\mathbf{D}$ is symmetric part of the velocity gradient defined by the relation $\mathbf{D} = \frac{1}{2}(\nabla \boldsymbol{u} + \nabla \boldsymbol{u}^T)$.

(b) **Viscoelastic fluids**
Maxwell model is the simplest model for viscoelastic fluid. In this case the stress tensor is computed from:

$$\mathbf{T} + \lambda_1 \frac{\delta \mathbf{T}}{\delta t} = 2\mu \mathbf{D} \tag{5}$$

where λ_1 has dimension of time and denotes the *relaxation time*. The symbol $\frac{\delta}{\delta t}$ represents upper convected derivative (see Eq. (9))

By combination of these two models the behaviour of mixture of viscous and viscoelastic fluids can be described. Such a model is called *Oldroyd-B model* and it has the form

$$\mathbf{T} + \lambda_1 \frac{\delta \mathbf{T}}{\delta t} = 2\mu \left(\mathbf{D} + \lambda_2 \frac{\delta \mathbf{D}}{\delta t} \right). \tag{6}$$

The parameters λ_1, λ_2 are *relaxation* and *retardation time*.

The stress tensor $\mathbf{T}$ is decomposed to the Newtonian part $\mathbf{T}_s$ and viscoelastic part $\mathbf{T}_e$ ($\mathbf{T} = \mathbf{T}_s + \mathbf{T}_e$) and

$$\mathbf{T}_s = 2\mu_s \mathbf{D}, \qquad \mathbf{T}_e + \lambda_1 \frac{\delta \mathbf{T}_e}{\delta t} = 2\mu_e \mathbf{D}, \tag{7}$$

where

$$\frac{\lambda_2}{\lambda_1} = \frac{\mu_s}{\mu_s + \mu_e}, \qquad \mu = \mu_s + \mu_e. \tag{8}$$

The *upper convected derivative* $\frac{\delta}{\delta t}$ is defined (for general tensor $\mathbf{M}$) by the relation (see [3])

$$\frac{\delta \mathbf{M}}{\delta t} = \frac{\partial \mathbf{M}}{\partial t} + (\boldsymbol{u}.\nabla)\mathbf{M} - (\mathbf{WM} - \mathbf{MW}) - (\mathbf{DM} + \mathbf{MD}) \tag{9}$$

where $\mathbf{D}$ is symmetric part of the velocity gradient

$$\mathbf{D} = \frac{1}{2}(\nabla \boldsymbol{u} + \nabla \boldsymbol{u}^T) = \frac{1}{2}\begin{pmatrix} 2u_x & u_y + v_x & u_z + w_x \\ u_y + v_x & 2v_y & v_z + w_y \\ w_x + u_z & w_y + v_z & 2w_z \end{pmatrix} \tag{10}$$

and $\mathbf{W}$ is antisymmetric part of the velocity gradient

$$\mathbf{W} = \frac{1}{2}(\nabla \boldsymbol{u} - \nabla \boldsymbol{u}^T) = \frac{1}{2}\begin{pmatrix} 0 & u_y - v_x & u_z - w_x \\ v_x - u_y & 0 & v_z - w_y \\ w_x - u_z & w_y - v_z & 0 \end{pmatrix}. \tag{11}$$

The governing system (1) and (2) of equations is completed by the equation for the viscoelastic part of the stress tensor

$$\frac{\partial \mathbf{T}_e}{\partial t} + (\boldsymbol{u}.\nabla)\mathbf{T}_e = \frac{2\mu_e}{\lambda_1}\mathbf{D} - \frac{1}{\lambda_1}\mathbf{T}_e + (\mathbf{WT}_e - \mathbf{T}_e\mathbf{W}) + (\mathbf{DT}_e + \mathbf{T}_e\mathbf{D}). \tag{12}$$

Both models could be generalized. In this case the viscosity μ is no more constant, but is defined by viscosity function according to the cross model (for more details see [12])

$$\mu(\dot{\gamma}) = \mu_\infty + \frac{\mu_0 - \mu_\infty}{(1 + (\lambda\dot{\gamma})^b)^a} \tag{13}$$

where

$$\dot{\gamma} = 2\sqrt{\frac{1}{2}\mathrm{tr}\,\mathbf{D}^2} \tag{14}$$

$\mu_0 = 1.6 \cdot 10^{-1}\, Pa \cdot s$ $\qquad$ $\mu_\infty = 3.6 \cdot 10^{-3}\, Pa \cdot \mathrm{s}$

$a = 1.23, b = 0.64$ $\qquad$ $\lambda = 8.2\, s.$

2 Numerical Solution

Numerical solution of the described models is based on cell-centered finite volume method using explicit Runge–Kutta time integration. The unsteady system of equations with steady boundary conditions is solved by finite volume method. Steady state solution is achieved for $t \to \infty$. In this case the artificial compressibility method can be applied. It means that the continuity equation is completed by the time derivative of the pressure in the form (for more details see e.g. [1,4,9]):

$$\frac{1}{\beta^2}\frac{\partial p}{\partial t} + \operatorname{div} \boldsymbol{u} = 0, \quad \beta \in \mathbb{R}^+. \tag{15}$$

The system of equations (including the modified continuity equation) could be rewritten in the vector form.

$$\tilde{R}_\beta W_t + F^c_x + G^c_y + H^c_z = F^v_x + G^v_y + H^v_z + S, \qquad \tilde{R}_\beta = \operatorname{diag}(\frac{1}{\beta^2}, 1, \cdots, 1) \tag{16}$$

where W is the vector of unknowns, F^c, G^c, H^c are inviscid fluxes, F^v, G^v, H^v are viscous fluxes defined as

$$W = \begin{pmatrix} p \\ u \\ v \\ w \\ t_1 \\ \vdots \\ t_6 \end{pmatrix}, \quad F^c = \begin{pmatrix} u \\ u^2 + p \\ uv \\ uw \\ ut_1 \\ \vdots \\ ut_6 \end{pmatrix}, \quad G^c = \begin{pmatrix} v \\ uv \\ v^2 + p \\ vw \\ vt_1 \\ \vdots \\ vt_6 \end{pmatrix}, \quad H^c = \begin{pmatrix} w \\ uw \\ vw \\ w^2 + p \\ wt_1 \\ \vdots \\ wt_6 \end{pmatrix}, \tag{17}$$

$$F^v = \begin{pmatrix} 0 \\ 2\mu(\dot{\gamma})u_x \\ \mu(\dot{\gamma})(u_y + v_x) \\ \mu(\dot{\gamma})(u_z + w_x) \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad G^v = \begin{pmatrix} 0 \\ \mu(\dot{\gamma})(u_y + v_x) \\ 2\mu(\dot{\gamma})v_y \\ \mu(\dot{\gamma})(v_z + w_y) \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad H^v = \begin{pmatrix} 0 \\ \mu(\dot{\gamma})(u_z + w_x) \\ \mu(\dot{\gamma})(v_z + w_y) \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \tag{18}$$

and the source term S is defined as

$$S = \begin{pmatrix} 0 \\ t_{1x} + t_{2y} + t_{3z} \\ t_{2x} + t_{4y} + t_{5z} \\ t_{3x} + t_{5y} + t_{6z} \\ 2\frac{\mu_e}{\lambda_1}u_x - \frac{t_1}{\lambda_1} + 2u_x t_1 + (u_y + v_x)t_2 + (u_z + w_x)t_3 \\ \frac{\mu_e}{\lambda_1}(u_y + v_x) - \frac{t_2}{\lambda_1} + u_y t_4 + u_z t_5 + u_y t_1 + w_y t_3 + u_x t_2 + v_y t_2 \\ \frac{\mu_e}{\lambda_1}(u_z + w_x) - \frac{t_3}{\lambda_1} + u_y t_5 + u_z t_6 + u_z t_1 + v_z t_2 + u_x t_3 + w_z t_3 \\ 2\frac{\mu_e}{\lambda_1}v_y - \frac{t_4}{\lambda_1} + (u_y + v_x)t_2 + 2v_y t_4 + (v_z + w_y)t_5 \\ \frac{\mu_e}{\lambda_1}(v_z + w_y) - \frac{t_5}{\lambda_1} + v_x t_3 + v_z t_6 + u_z t_2 + v_z t_4 + v_y t_5 + w_z t_5 \\ 2\frac{\mu_e}{\lambda_1}w_z - \frac{t_6}{\lambda_1} + (u_z + w_x)t_3 + (v_z + w_y)t_5 + 2w_z t_6 \end{pmatrix} \tag{19}$$

The following special parameters settings related to four specific models will be used in our numerical simulation:

Newtonian	$\mu(\dot{\gamma}) = \mu_s = const.$	$\mathbf{T}_e \equiv 0$
Generalized Newtonian	$\mu(\dot{\gamma})$	$\mathbf{T}_e \equiv 0$
Oldroyd-B	$\mu(\dot{\gamma}) = \mu_s = const.$	$\mathbf{T}_e$
Generalized Oldroyd-B	$\mu(\dot{\gamma})$	$\mathbf{T}_e$

The Eq. (16) is discretized in space by the cell-centered finite volume method (see [6]) and the arising system of ODEs is integrated in time by the explicit multistage Runge–Kutta scheme (see [7, 11, 12]):

$$\begin{aligned} W_i^n &= W_i^{(0)} \\ W_i^{(s)} &= W_i^{(0)} - \alpha_{s-1}\Delta t \mathscr{R}(W)_i^{(s-1)} \\ W_i^{n+1} &= W_i^{(M)} \qquad\qquad s = 1, \ldots, M, \end{aligned} \tag{20}$$

where $M = 3, \alpha_0 = \alpha_1 = 0.5, \alpha_2 = 1.0$, the steady residual $\mathscr{R}(W)_i$ is defined by finite volume method as

$$\mathscr{R}(W)_i = \frac{1}{\sigma_i}\sum_{k=1}^{6}\left[\left(\overline{F}_k^c - \overline{F}_k^v\right)\Delta Sx_k + \left(\overline{G}_k^c - \overline{G}_k^v\right)\Delta Sy_k + \left(\overline{H}_k^c - \overline{H}_k^v\right)\Delta Sz_k\right] - \overline{S}, \tag{21}$$

where σ_i is the volume of the cell, $\sigma_i = \int\int_{C_i} dx\, dy\, dz$. The symbols $\overline{F}_k^c, \overline{G}_k^c, \overline{H}_k^c$ and $\overline{F}_k^v, \overline{G}_k^v, \overline{H}_k^v$ denote the numerical approximation of the inviscid and viscous fluxes. The symbol $\overline{S}$ represents the numerical approximation of the source term

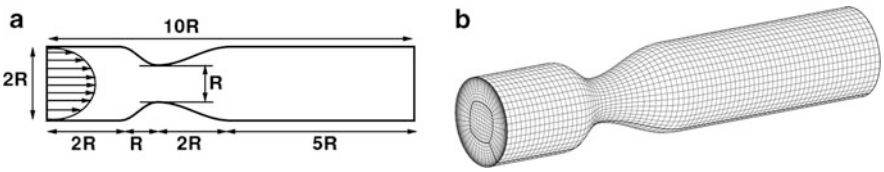


Fig. 1 Structure of the computational domain. (**a**) Newtonian. (**b**) Generalized Newtonian

with central approximation of derivatives. Sx_k, Sy_k, Sz_k are the projection of surface of k-th face of cell C_i into the plane which is ortogonal to axis x, y, z. The numerical approximation of the inviscid fluxes are computed as an arithmetic average of the inviscid numerical fluxes of two neighbouring finite volume cells. The viscous fluxes contain the velocity derivatives and their numerical approximation are computed around the Gauss theorem. For more details in 2D case see [8].

The multistage Runge–Kutta scheme (21) is conditionally stable. The time step is chosen to satisfy the CFL coditions (for orthogonal grid)

$$\Delta t \leq \frac{CFL}{\frac{\rho_A}{Sx} + \frac{\rho_B}{Sy} + \frac{\rho_C}{Sz} + \frac{\mu_s}{\rho}\left(\frac{1}{Sx^2} + \frac{1}{Sy^2} + \frac{1}{Sz^2}\right)} \tag{22}$$

where ρ_A, ρ_B, ρ_C are spectral radius of Jacobi matrices of inviscid fluxes F, G, H.

2.1 Boundary Conditions

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 is used and for a pressure and the stress tensor Neumann boundary condition is used. At the outlet the pressure value is given and for the velocity vector and the stress tensor Neumann boundary condition is used. The homogeneous Dirichlet boundary condition for the velocity vector is used on the wall. For the pressure and stress tensor Neumann boundary condition is considered.

3 Numerical Results

This section deals with the comparison of the numerical results of Newtonian and Oldroyd-B fluids. Numerical tests are performed in an idealized stenosed vessel. The stenosed vessel is assumed to be three-dimensional with circular cross-section (see Fig. 1). Figure 4 shows the shape of the tested domain. The computational

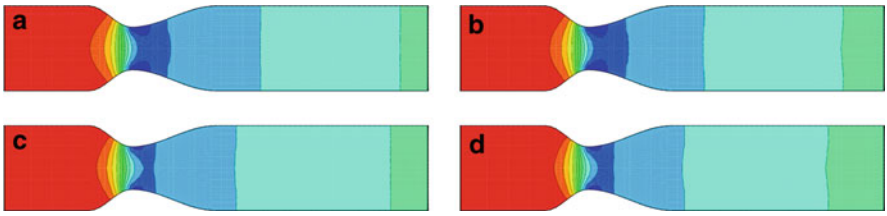


Fig. 2 Pressure distribution for generalized Oldroyd-B fluids. (**a**) Newtonian. (**b**) Generalized Newtonian. (**c**) Oldroyd-B. (**d**) Generalized Oldroyd-B

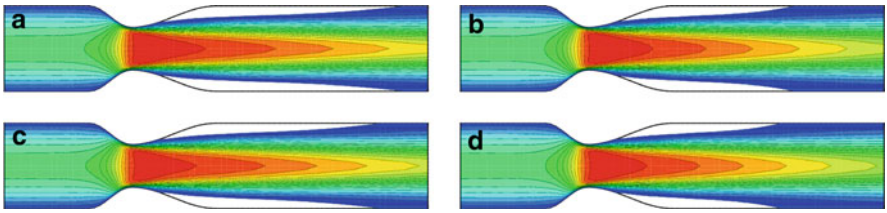


Fig. 3 Axial velocity isolines for generalized Oldroyd-B fluids. (**a**) Newtonian. (**b**) Generalized Newtonian. (**c**) Oldroyd-B. (**d**) Generalized Oldroyd-B

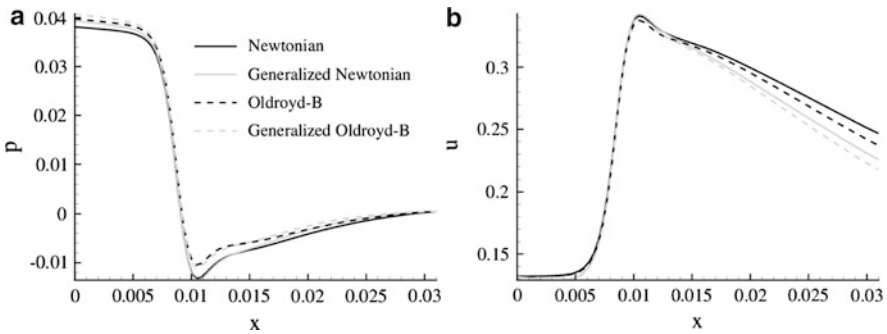


Fig. 4 Pressure and axial velocity distribution along the central axis of the channel. (**a**) Pressure. (**b**) Axial velocity

domain is discretized using a structured, wall fitted mesh with hexahedral cells and uniform axial cell spacing. The outer mesh block has $40 \times 16 \times 80$ cells, while the central mesh block has $10 \times 10 \times 80$ finite volume cells.

The following model parameters are:

$$\mu_e = 4.0 \cdot 10^{-4}\, Pa \cdot s \qquad \mu_s = 3.6 \cdot 10^{-3}\, Pa \cdot s$$
$$\lambda_1 = 0.06\, s \qquad \lambda_2 = 0.054\, s$$
$$U_0 = 0.0615\, m \cdot s^{-1} \qquad L_0 = 2R = 0.0062\,\text{m}$$
$$\mu_0 = \mu = \mu_s + \mu_e \qquad \rho = 1{,}050\, kg \cdot m^{-3}$$

In Figs. 2 and 3 the comparison of the pressure distributions and the axial velocity isolines is presented. To emphasize the flow separation behind the stenosis the regions of reversal flow (with respect to axial direction) are marked with white color.

Pressure and velocity distribution along the axis for both tested fluids models is shown in Fig. 4. By simple observation one can conclude that the main effect of the Oldroyd-B fluids behavior is visible mainly in the recirculation zone. For obtaining the steady solution for all the tested cases 100,000 time steps are needed.

4 Conclusions

Newtonian and Oldroyd-B models with their generalized modification have been considered for numerical simulation of fluids flow in the idealized axisymmetric stenosis. The cell-centered finite volume solver for incompressible laminar viscous and viscoelastic fluids flow has been described. For time integration the explicit Runge–Kutta method was considered. The numerical results obtained by this method are presented. The differences between these tested fluids are given mainly in the separation region. These results clearly show that for shear-thinning flows the recirculation zone becomes shorter. This could be explained by the specific choice of the characteristic viscosity μ_∞ for the reference Newtonian and (non-generalized) Oldroyd-B solution.

Acknowledgements This work was partly supported by the grant GACR P201/11/1304 and GACR 201/09/0917.

References

1. P. Louda, K. Kozel, J. Příhoda, L. Beneš, T. Kopáček. 2011. Numerical solution of incompressible flow through branched channels, *Computers & Fluids*, 46:318–324
2. T. Bodnar, A. Sequeira, M. Prosi. 2010. On the shear-thinning and viscoelastic effects of blood flow under various flow rates, *Applied Mathematics and Computation*
3. T. Bodnar, A. Sequeira. 2010. Numerical study of the significance of the non-Newtonian nature of blood in steady flow through s stenosed vessel, *Advances in Mathematical Fluid Mechanics*, 83–104.
4. A.J. Chorin. 1967. A numerical method for solving incompressible viscous flow problem, *Journal of Computational Physics* 135:118–125
5. R. Dvořák, K. Kozel. 1996. *Mathematical Modelling in Aerodynamics (in Czech)*. Prague Czech Republic: CTU.
6. R. LeVeque. 2004. *Finite-volume methods for hyperbolic problems*, Cambridge University Press
7. R. Keslerová, K. Kozel. 2010. Numerical modelling of incompressible flows for Newtonian and non-Newtonian fluids, *Mathematics and Computers in Simulation* 80:1783–1794
8. R. Keslerová, K. Kozel. 2011. Numerical solution of laminar incompressible generalized Newtonian fluids flow, *Applied Mathematics and Computation*, 217:5125–5133

9. R. Keslerová, K. Kozel. 2008. Numerical simulations of incompressible laminar flow for Newtonian and non-Newtonian fluids, In *Numerical Mathematics and Advanced Applications, ENUMATH 2007*, ed. Springer-Verlag Berlin
10. A.M. Robertson, A. Sequeira, M.V. Kameneva. 2008. *Hemorheology*. Switzerland: Birkhäuser Verlag Basel.
11. A. Jameson, W. Schmidt, E. Turkel. 1981. Numerical solution of the Euler equations by finite volume methods using Runge-Kutta time-stepping schemes, *AIAA 14th Fluid and Plasma Dynamic Conference California*
12. J. Vimmr, A. Jonášová. Non-Newtonian effects of blood flow in complete coronary and femoral bypasses, *Mathematics and Computers in Simulation*, 80:1324–1336

Numerical Simulations of Turbulent 3D Flow in Channel Junction

P. Louda, K. Kozel, J. Příhoda, and L. Beneš

Abstract The work deals with numerical 3D simulations of incompressible turbulent flow in channel junction with one inlet and two outlets. The complex flow in the junction includes separation, impingement and secondary flow. The mathematical model is based on unsteady Reynolds averaged Navier-Stokes equations (URANS) with an explicit algebraic Reynolds stress turbulence model (EARSM). The solution method uses dual time artificial compressibility scheme with upwind finite volume discretization. Some methods of ensuring prescribed flow-rate distribution are discussed and tested. The results are compared with PIV measurement.

1 Introduction

In this work, the turbulent incompressible flow through a junction of three perpendicular channels is modelled numerically. This case is often encountered in engineering applications and pressure losses are of interest. The T-shape geometry of channel moreover leads to different flow phenomena according to how the inlets and outlets are arranged. Two examples are given in Fig. 1. The case shown on the left part of the figure has been considered by the authors in [5]. The other case is considered here. The inlet flow through the branch creates impingement in the main channel, then recirculation zones above and secondary flows as well. As it shows, for the conditions prescribed, the flow becomes unsteady. Nevertheless, the Reynolds averaged model of the turbulence was used, which is interpreted as unsteady RANS. The mathematical model is described in next section. The other

P. Louda (✉) · K. Kozel · L. Beneš
Czech Technical University in Prague, Prague, Czech Republic
e-mail: petr.louda@fs.cvut.cz; karel.kozel@fs.cvut.cz; ludek.benes@fs.cvut.cz

J. Příhoda
Institute of Thermomechanics, Czech Academy of Sciences, Prague, Czech Republic
e-mail: prihoda@it.cas.cz

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_57,

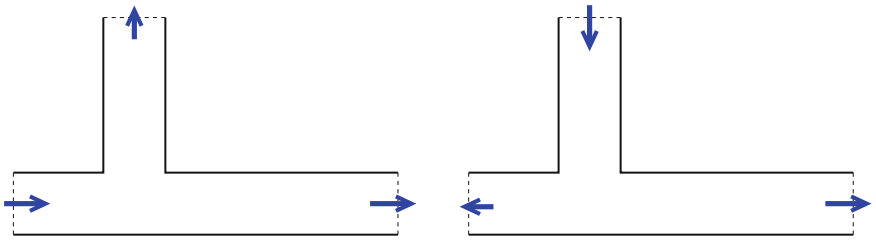


Fig. 1 Two examples of flow arrangement

aspect of the simulations is the need to fulfill prescribed flow rates in each tube. This inverse problem is solved iteratively by adjusting static pressure in the inlet and the outlets. A comparison with measured data is presented as well.

2 Mathematical Model and Solution Method

The mathematical model is based on Reynolds averaged Navier-Stokes equations in Cartesian coordinates for 3D turbulent incompressible flow. For unsteady simulation it formally becomes URANS (unsteady RANS) approach. The physical meaning is maintained if the simulated unsteadiness is far enough from turbulent scales. The averaged equations formally differ from Navier-Stokes equations by additional momentum transport expressed by Reynolds stress tensor. In this work, the explicit algebraic Reynolds stress model (EARSM) [2,6] is used. The model requires solving a system of k-ω equations for turbulent energy k and specific dissipation rate ω.

The governing URANS equations are solved by artificial compressibility method extended for unsteady cases by use of dual time. In the simplest form of the artificial compressibility method [1], only the continuity equation is modified by adding pressure time derivative

$$\frac{1}{\beta^2}\frac{\partial(p/\rho)}{\partial t} + \frac{\partial u_1}{\partial x_1} + \frac{\partial u_2}{\partial x_2} + \frac{\partial u_3}{\partial x_3} = 0, \ \rho = const \tag{1}$$

where β is a positive artificial compressibility parameter. The system including the modified continuity equation and the momentum equations can be written for a domain of solution Ω in the following form

$$\begin{aligned} &\Gamma\frac{\partial W}{\partial t} + Rez(W) = 0, \\ &\Gamma = diag[\beta^{-2}, 1, 1, 1, 1, 1], \ W = col[p/\rho, u_1, u_2, u_3, k, \omega], \\ &(x_1, x_2, x_3) \in \Omega, \quad t \in (0, \infty) \end{aligned} \tag{2}$$

where W is vector of unknown kinematic pressure and velocity components and turbulent scales, and steady residual $Rez(W)$ contains all inviscid and viscous terms. However, the divergence free velocity field is not achieved before steady state, $\partial p/\partial t = 0$. To satisfy continuity at each time level, dual (iterative) time τ between two time grid levels (t_n, t_{n+1}) is introduced in the system

$$\Gamma \frac{\partial W}{\partial \tau} + Rez^{uns}(W) = 0, \quad Rez^{uns}(W) = R\frac{\partial W}{\partial t} + Rez(W), \; R = diag[0,1,1,1], \quad (3)$$
$$(x_1, x_2, x_3) \in \Omega, \quad t \in < t_n, t_{n+1} >, \quad \tau \in (0, \infty)$$

for $\tau \in (0, \infty)$. Numerical solution in τ is realized iteratively for $\tau = \tau_0 = t_n, \; \tau_1, \; \tau_2, \ldots$ and iterative process in τ is considered in such a way that for $\tau \to \infty$ we achieve $W^n \to W^{n+1}$.

The time discretization scheme considered here is implicit both for physical and artificial time. For the dual time method, the scheme is combination of the backward Euler method for artificial time (superscript ν) and three-layer scheme for physical time (superscript n)

$$\Gamma \frac{W^{\nu+1}_{i,j,k} - W^{\nu}_{i,j,k}}{\Delta\tau} + R\frac{3W^{\nu+1}_{i,j,k} - 4W^{n}_{i,j,k} + W^{n-1}_{i,j,k}}{2\Delta t} + Rez(W)^{\nu+1}_{i,j,k} = 0. \quad (4)$$

where physical time step Δt is chosen according to the solved problem. The $\Delta\tau$ for explicit scheme is a suitable function of Δt and spatial grid steps which fulfills the condition that iterative process in τ is stable. In our case the scheme in τ is implicit and $\Delta\tau$ is not limited. It is chosen to be very large $\approx 10^7 \Delta t$ in order to achieve W^{n+1} quickly in eight to ten iteration steps.

The steady residuals are computed by a cell-centered finite volume method with hexahedral finite volumes with third order upwind approximation for convective terms. The method is described in more detail in [4].

2.1 Boundary Conditions

The 3D computational domain has boundaries consisting of solid walls, one inlet and two outlets. On the wall, velocity is zero. In the inlet, velocity and turbulence model variables k, ω are prescribed according to developed channel flow. In the outlets, velocity and turbulence model variables are linearly extrapolated.

The flow conditions are prescribed by flow-rates in each branch according to measurement. Since the flow contains large scale unsteadiness, there is some uncertainty in specifying constant flow rate in both outlets. Nevertheless this is the selected approach.

The target flow rates are achieved using two conditions for pressure in the outlets:

(α) "do-nothing"-like condition

$$\mu \frac{\partial u}{\partial n} - p = p_{ref} \tag{5}$$

where u is velocity component normal to the outlet plane, μ dynamic viscosity, p pressure and p_{ref} an arbitrary constant

(β) Correction for target flow rate U_{bt}

$$\frac{\partial p}{\partial n} = -\frac{U_{bt} - U_b}{\Delta t} \tag{6}$$

where U_b is flow rate at time t_n and $\Delta t = t_{n+1} - t_n$

The condition α can be used in one outlet only and needs to be combined with e.g. condition β. The condition β can be used in both outlets. Any of these three combinations worked comparably well in the simulated cases.

3 Numerical Results

The simulations have been carried out on finite volume mesh consisting of 19 blocks with total of 790,000 finite volumes. The Reynolds number based on the inlet channel diameter D and bulk inlet velocity was $Re = 140{,}000$. In the experiment it corresponds to flow rate of 5.5 l/s. The main channel with two outlets has the same diameter D and the flow is divided in the outlets according to prescribed ratio 50:50 or 20:80. The orientation of coordinates is shown in Fig. 2.

The Fig. 3 shows isolines time averaged velocity in the symmetry plane as well as an example of instantaneous velocity. There seems to be not much unsteadiness except in the recirculation zone in the left branch with smaller flow-rate. The next Fig. 3 shows isolines of kinetic energy of resolved velocity fluctuations in comparison with modelled turbulent energy k. One can see that the maximum of resolved unsteadiness is the recirculation zone in the right branch and the turbulent energy reaches the level of modelled turbulent energy (Fig. 4).

The next Fig. 5 shows comparison of computational results with PIV measurement by Kotek et al. [3] in terms of isolines of velocity and velocity vectors in the symmetry plane. The computed position of the stagnation point is shifted more to the left. The most differences are observed in the recirculation zones.

The velocity profiles in six positions marked in the Fig. 2 are shown in next figures. First, the inlet channel profiles, Fig. 6 suggest that the simulation may have higher flow-rate. The difference could be explained by three-dimensionality of the

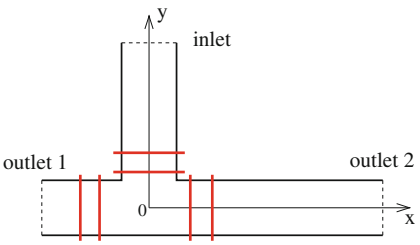


Fig. 2 Coordinates and six cuts where comparison with measurement are made

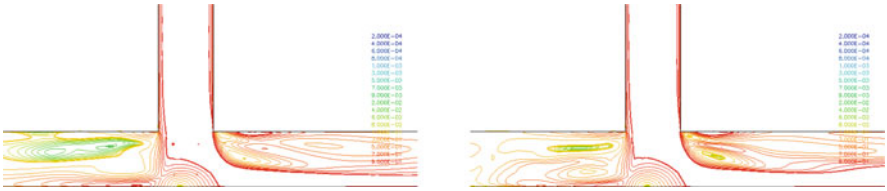


Fig. 3 Isolines of velocity in the symmetry plane, division ratio 20:80. *Left*: time averaged, *right*: instantaneous

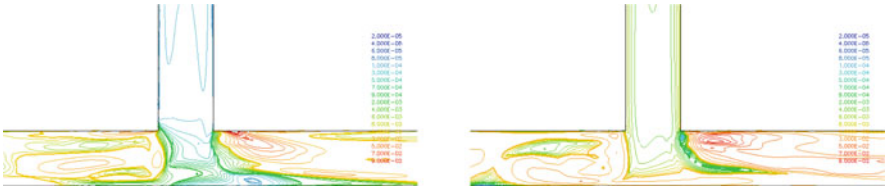


Fig. 4 Isolines of turbulent energy in the symmetry plane, division ratio 20:80. *Left*: resolved energy, *right*: modelled energy k

velocity field of which we do not have experimental data. The velocities in the cross-sections of the main channel are shown in Fig. 7. Finally the Fig. 8 show velocity on axes of the channels.

For the case with flow-rate division 50:50, the Fig. 9 shows pressure along the channels in terms of time average and one example of instantaneous pressure. The three variants denote the combinations conditions for pressure in the outlets according to description in Sect. 2.1:

- var 1: condition β in the left outlet, condition α in the right outlet,
- var 2: condition α in the left outlet, condition β in the right outlet,
- var 3: conditions β in both outlets,

whereas in the inlet condition β was always used. It can be seen that the distribution in main channel is practically symmetric as expected, although the geometry was not symmetric – the left branch was noticeably shorter than the right one.

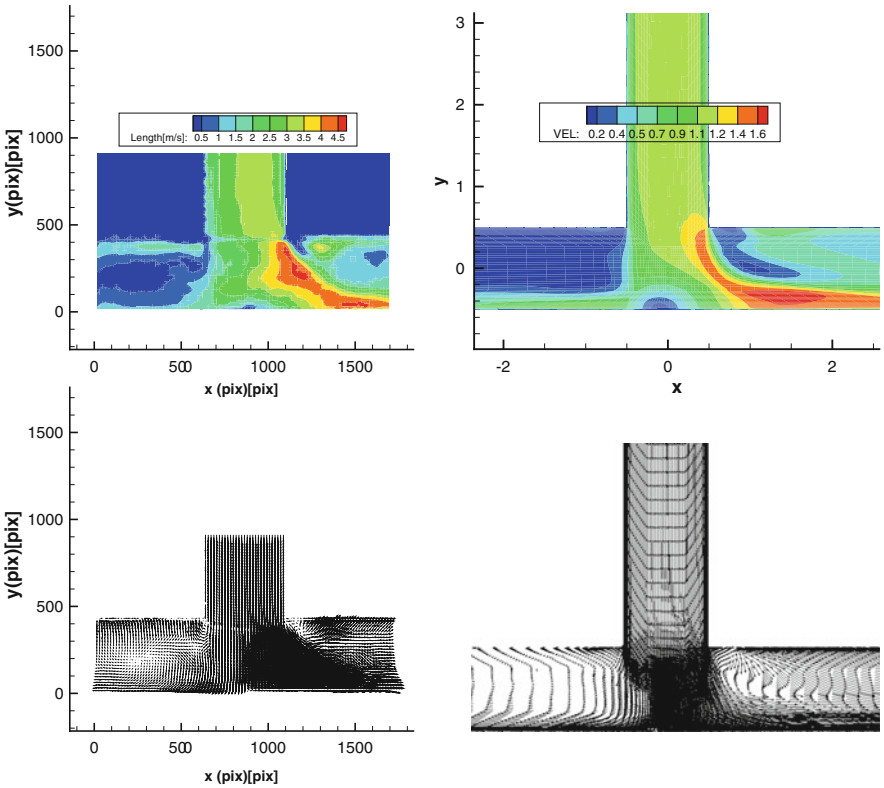


Fig. 5 Comparison of simulation (*left*) with PIV experiment (*right*), division 20:80

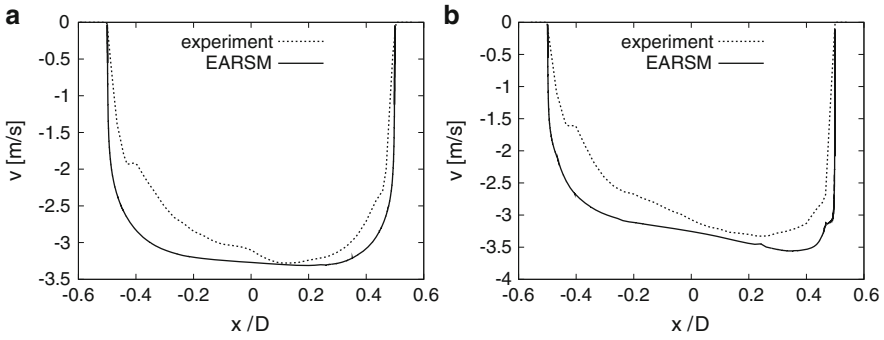


Fig. 6 Comparison with PIV experiment, velocity in the inlet branch, averaged flow at $Re = 140{,}000$, flow-rate division 20:80, total flow-rate 5.5 l/s. (**a**) $y = D$. (**b**) $y = 0.62D$

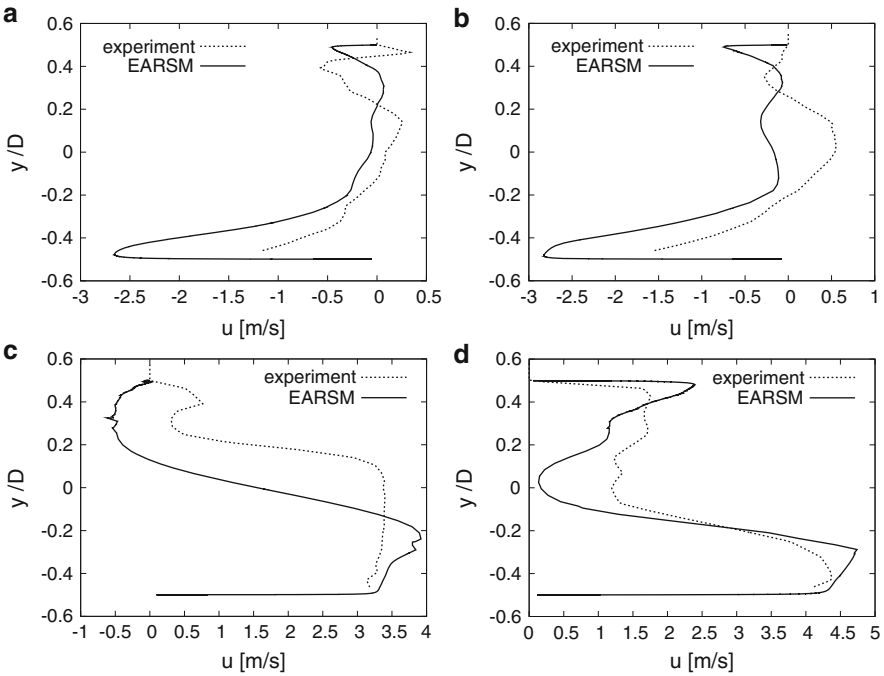


Fig. 7 Comparison with PIV experiment, velocity in the main channel, averaged flow at $Re = 140{,}000$, flow-rate division 20:80, total flow-rate 5.5 l/s. (**a**) $x = -1.25D$. (**b**) $x = -0.8D$. (**c**) $x = 0.75D$. (**d**) $x = 1.20D$

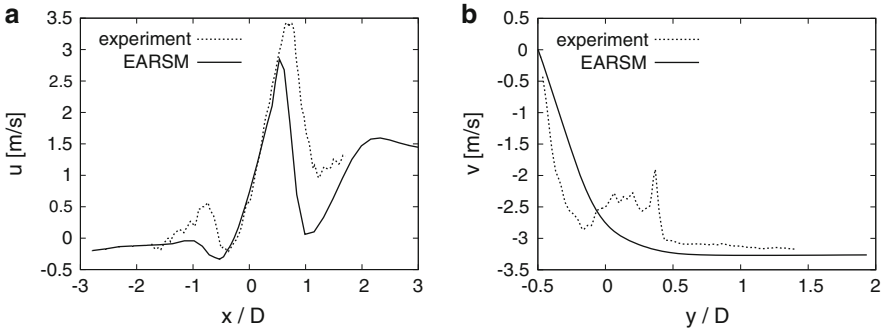


Fig. 8 Comparison with PIV experiment, velocity on channel axes, averaged flow at $Re = 140{,}000$, flow-rate division 20:80, total flow-rate 5.5 l/s. (**a**) Main channel. (**b**) Inlet branch

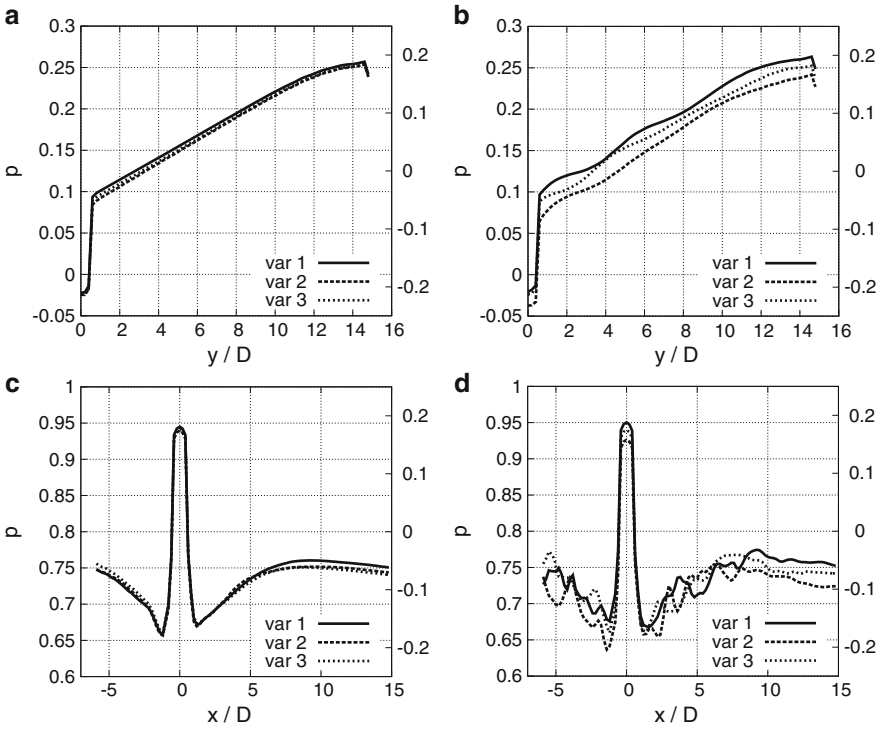


Fig. 9 Pressure averaged over cross-section area, division 50:50, round channel, $Re = 140{,}000$. (**a**) Time averaged pressure in the inlet branch. (**b**) Instantaneous pressure in the inlet branch. (**c**) Time averaged pressure in the main channel. (**d**) Instantaneous pressure in the main channel

4 Conclusions

The work presents URANS simulation of flow in junction of three pipes of equal cross-section. Mathematical model based on artificial compressibility method with dual time is presented. The conditions to fulfill prescribed flow-rate distribution were discussed and tested. The results are compared with PIV measurement. The main qualitative parameters of velocity field were captured well, with some discrepancies in the quantitative agreement. This can be explained by rough, but the only feasible, modeling approach, i.e. URANS.

Acknowledgements The work was partially supported by grants No. P101/10/1230, 103/09/0977 and 101/09/1539 of the Czech Science Foundation.

References

1. A. J. Chorin. A numerical method for solving incompressible viscous flow problems. *J. of Computational Physics*, 2(1):12–26, 1967.
2. A. Hellsten. New advanced k-ω turbulence model for high-lift aerodynamics. *AIAA J.*, 43:1857–1869, 2005.
3. M. Kotek, D. Jašíková, and V. Kopecký. Experimental study of the flow field in T-junction model using piv method. In *17th Int. Conf. Engineering Mechanics 2011*, Svratka, Czech Republic, 2011.
4. P. Louda, K. Kozel, and J. Příhoda. Numerical solution of 2D and 3D viscous incompressible steady and unsteady flows using artificial compressibility method. *Int. J. for Numerical Methods in Fluids*, 56:1399–1407, 2008.
5. P. Louda, K. Kozel, J. Příhoda, L. Beneš, and T. Kopáček. Numerical solution of incompressible flow through branched channels. *Computers and Fluids*, 46:318–324, 2011.
6. S. Wallin. *Engineering turbulence modeling for CFD with a focus on explicit algebraic Reynolds stress models*. PhD thesis, Royal Institute of Technology, Stockholm, 2000.

Weak Formulation of the Problem of Modelling the Steady Flow of a Viscous Incompressible Liquid Through a Rotating Radial Blade Machine

T. Neustupa

Abstract The paper presents the mathematical model of a two dimensional steady viscous incompressible flow through a rotating radial blade machine. The flow is described and studied in the rotating frame. The paper provides the classical and weak formulation of the corresponding boundary value problem. The boundary condition on the outflow is the so called "natural" boundary condition, with the additional nonlinear term proposed by Bruneau and Fabrie (Math Model Numer Anal 30(7):815–840, 1996), and a new term arising from the rotation of the machine. The existence of a weak solution is proved.

1 Introduction and the Geometry of the Problem

One of types of turbines, often used in mechanical engineering, is the so called Kaplan turbine. This kind of turbines has a wide field of applications, e.g. in metallurgy, energetics, etc. Mathematics often models flows through turbines by means the infinite two–dimensional cascade of profiles. This approach has several advantages, e.g. it leads to the spatial periodic flow and thus enables one to reduce the problem from an originally unbounded domain to just one spatial period which is a bounded domain. The mathematical model of Kaplan's turbine is geometrically different. The considered domain is bounded from the beginning and the parts of the boundary, where the fluid enters or leaves the domain, are concentric circles. The outside circle represents the inflow and the inside circle is the outflow. The blades of the turbine are regularly placed between the inflow and the outflow, see example on Fig. 1.

T. Neustupa (✉)
Faculty of Mechanical Engineering, Czech Technical University Prague, Karlovo nám. 13, 121 35, Prague, Czech Republic
e-mail: tneu@centrum.cz

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_58,

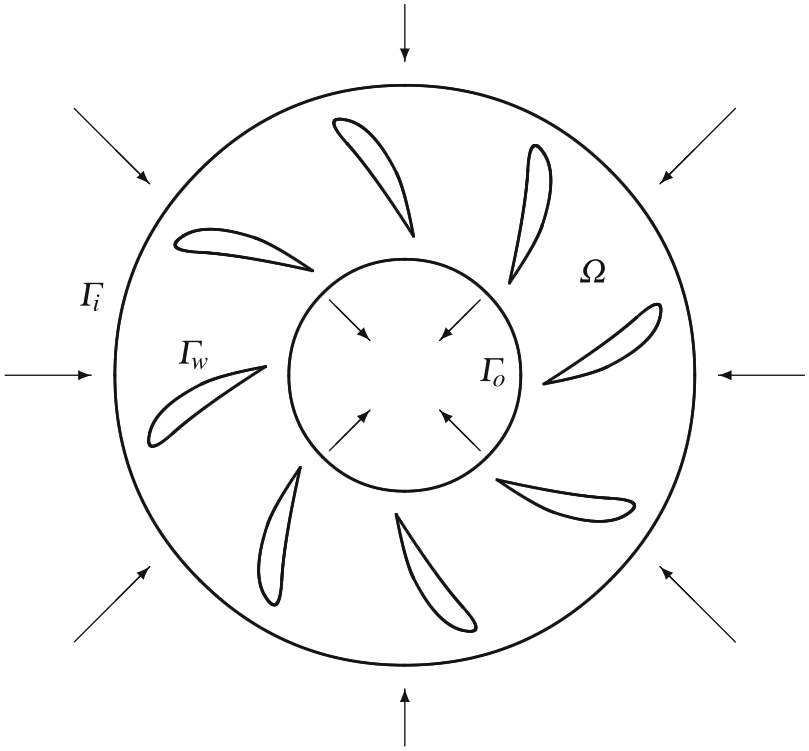


Fig. 1 Radial cascade of profiles, domain Ω

The flow of a viscous incompressible fluid through the rotating turbine is described by the Navier–Stokes equations, appropriately transformed to the rotating coordinates. The form of these equations is well known, see e.g. [2, 3]. We use the two–dimensional version of the transformed equation. The considered boundary conditions on the inflow and on the turbine blades are of the Dirichlet type – inhomogeneous on the inflow and copying the rotational velocity on the blades.

Dirichlet's boundary condition, however, is not convenient on the outflow because the velocity profile on the outflow is not known in advance. The question of an appropriate boundary condition on the outflow is similar to the situation at the outflow from a channel. The natural boundary condition (also often called the "do nothing" boundary condition), used e.g. in [4] and [5], does not exclude the possibility of a backward flow, which might bring an uncontrollable amount of the kinetic energy to the flow field. This means that one cannot formally derive the usual energy inequality, which in other situations represents the basic a priori estimate of a solution. This is why we use the modified natural boundary condition. It contains the additional nonlinear term, which enables us to control the kinetic energy of the backward flow. The same modification was proposed in [1] in connection with the flow in a channel.

The main result of this paper is the theorem on the existence of a weak solution. The studied boundary value problem differs from other mostly used formulations by additional terms in the Navier–Stokes equations which come from the transformation to the rotating coordinates, by the special geometry of the domain and by the choice of the boundary conditions. The used weak formulation can be considered as a starting point for numerical analysis and numerical solution, based on the method of finite elements.

2 Classical Formulation of the Boundary Value Problem

2.1 Geometry of the Domain

Domain Ω_t is a 2D annulus between two concentric circles, without the blades of the turbine – see Fig. 1. The outside circle Γ_i is the inflow part of the boundary and the inside circle Γ_o is the supposed outflow from the domain. The subscript t expresses that Ω_t is time–varying due to the rotation of the turbine.

Remark 1. Kaplan's turbine is in fact a three dimensional object. While the flow in domain Ω around the blades can be considered to be two–dimensional, the flow inside the smaller circle rapidly changes the direction into the third dimension and it cannot be therefore modelled as a 2D flow. However, the flow inside the smaller circle is not involved in the model studied in this paper.

2.2 Transformation to the Rotating Coordinates

We suppose that the considered Kaplan turbine rotates about its center with the angular velocity ω. Thus, from the point of view of an outside observer, domain Ω_t is time–varying. To obtain a problem in a time–independent domain, we transform the mathematical problem to the coordinates $\mathbf{x}' = (x_1', x_2')$, whose origin is in the center of the turbine and which are connected with the rotating turbine. If the original Cartesian coordinate system is $\mathbf{x} = (x_1, x_2)$ (with the origin in the center of the turbine) then the transformation is

$$\mathbf{x}' \;=\; \begin{pmatrix} x_1' \\ x_2' \end{pmatrix} \;=\; R(t) \begin{pmatrix} x_1 \\ x_2 \end{pmatrix},$$

where

$$R(t) \;=\; \begin{pmatrix} \cos(\omega t), & \sin(\omega t) \\ -\sin(\omega t), & \cos(\omega t) \end{pmatrix}.$$

For $\omega > 0$ is the rotation in clockwise direction.

If we denote by $\mathbf{u} = (u_1, u_2)$ (respectively $\mathbf{u}' = (u_1', u_2')$) the velocity in the fixed coordinate system $\mathbf{x}$ (respectively the rotating coordinate system $\mathbf{x}'$) then the relations between $\mathbf{u}$ and $\mathbf{u}'$ are

$$\mathbf{u}'(t, \mathbf{x}') = R(t)\,\mathbf{u}(t, \mathbf{x}) = R(t)\,\mathbf{u}\big(t, R^{-1}(t)\mathbf{x}'\big),$$

$$\mathbf{u}(t, \mathbf{x}) = R^{-1}(t)\,\mathbf{u}'(t, \mathbf{x}') = R^{-1}(t)\,\mathbf{u}'\big(t, R(t)\mathbf{x}\big).$$

It can be calculated that the Navier–Stokes equation

$$\partial_t \mathbf{u} + (\mathbf{u}\cdot\nabla)\mathbf{u} = \mathbf{f} - \nabla p + \nu\,\Delta\mathbf{u},$$

transformed to the rotating frame, has the form

$$\partial_t \mathbf{u}' + (\mathbf{u}'\cdot\nabla')\mathbf{u}' + \omega\,\mathbf{u}'^{\perp} - \omega\,(\mathbf{x}'^{\perp}\cdot\nabla')\mathbf{u}' = \mathbf{f} - \nabla' p + \nu\,\Delta'\mathbf{u}',$$

see e.g. [2, 3]. Here, we denote by ∇', respectively Δ', the operators acting in the coordinate system $\mathbf{x}'$. The superscript $\perp$ means the perpendicular vector, i.e. $\mathbf{x}'^{\perp} = (x_1', x_2')^{\perp} = (-x_2', x_1')$. The condition of incompressibility $\operatorname{div}\mathbf{u} = 0$ is transformed to $\operatorname{div}'\mathbf{u}' = 0$. These equations are considered in domain Ω', which is the set of $\mathbf{x}'$ of the form $\mathbf{x}' = R(t)\mathbf{x}$ for $\mathbf{x} \in \Omega_t$. Domain Ω' is time–independent.

Further, we deal only with the quantities in the rotating frame $\mathbf{x}'$. In order to simplify the notation, we omit writing the primes. Moreover, since we are interested in the steady problem (in the rotating frame), we omit the time derivative of velocity. Thus, the considered system of equations becomes

$$(\mathbf{u}\cdot\nabla)\mathbf{u} + \omega\,\mathbf{u}^{\perp} - \omega\,(\mathbf{x}^{\perp}\cdot\nabla)\mathbf{u} = \mathbf{f} - \nabla p + \nu\,\Delta\mathbf{u} \qquad \text{in } \Omega, \tag{1}$$

$$\operatorname{div}\mathbf{u} = 0 \qquad \text{in } \Omega. \tag{2}$$

The velocity on the surface of the blades should copy, due to the no–slip condition, the local velocity of the blade. It equals $\omega\mathbf{x}^{\perp}$. Hence we get the boundary condition

$$\mathbf{u}\,|_{\Gamma_w} = \omega\mathbf{x}^{\perp}. \tag{3}$$

We assume that the velocity profile on the inflow part of boundary Γ_i is known. The given velocity on Γ_i becomes, after the transformation, $\mathbf{g} + \omega\mathbf{x}^{\perp}$. Thus, we get the boundary condition

$$\mathbf{u}\,|_{\Gamma_i} = \mathbf{g} + \omega\mathbf{x}^{\perp}. \tag{4}$$

We apply the nonlinear "natural" boundary condition on the outflow part of boundary Γ_o in the form

$$-\nu\,\frac{\partial\mathbf{u}}{\partial\mathbf{n}} + p\mathbf{n} - \frac{1}{2}\,(\mathbf{u}\cdot\mathbf{n})^{-}\,\mathbf{u} + \frac{1}{2}\,(\mathbf{u}\cdot\mathbf{n})^{-}\,\omega\mathbf{x}^{\perp} = \mathbf{0} \qquad \text{on } \Gamma_o. \tag{5}$$

Here, $\mathbf{n}$ denotes the unit outer normal vector to $\partial\Omega$. The minus–superscript denotes the negative part. (I.e. $a^{-} = (|a| - a)/2$.) Condition (5) is a modification of the

boundary condition $-\nu\,(\partial\mathbf{u}/\partial\mathbf{n}) + p\,\mathbf{n} = \mathbf{0}$ on Γ_o, used e.g. in [4]. It follows "naturally" from the weak formulation of the boundary value problem in a fixed (i.e. non–rotating) domain. The artificial term $\frac{1}{2}\,(\mathbf{u}\cdot\mathbf{n})^-\,\mathbf{u}$ acts only in the case of backward flows on Γ_o and it enables us to derive an estimate, which we later need in order to prove the existence and convergence of Galerkin approximations. The natural boundary condition in the form

$$-\nu\,\frac{\partial\mathbf{u}}{\partial\mathbf{n}} + p\,\mathbf{n} - \frac{1}{2}\,(\mathbf{u}\cdot\mathbf{n})^-\,\mathbf{u} = \mathbf{0} \qquad \text{on } \Gamma_o$$

was suggested and used in [1]. The term $\frac{1}{2}\,(\mathbf{u}\cdot\mathbf{n})^-\,\omega\,\mathbf{x}^\perp$ on the left hand side of (5) is added due to the rotation of the turbine. It enables us to formulate a relatively simple corresponding "weak problem" (see Sect. 3.1). On the other hand, condition (5) "naturally" follows from the week formulation of the problem in the rotating frame, assuming that the weak solution is sufficiently smooth and applying the backward integration by parts.

We further need an appropriate extension of function $\mathbf{g}$ from Γ_i to Ω. The next lemma provides the information on the existence of such extension.

Lemma 1. *Let $\mathbf{g} \in W^{1/2,2}(\partial\Omega)^2$. There exists a constant $c_1 > 0$ and a divergence–free extension $\mathbf{g}^* \in W^{1,2}(\Omega)^2$ of function $\mathbf{g}$ from Γ_i into Ω such that $\mathbf{g}^* = 0$ on Γ_w and*

$$\|\mathbf{g}^*\|_{1,2} \le c_1\,\|\mathbf{g}\|_{1/2,2;\,\Gamma_i} \tag{6}$$

where constant c_1 is independent of $\mathbf{g}$.

Here, we denote by $\|\,.\,\|_{1/2,2;\,\Gamma_i}$ the norm in the the Sobolev–Slobodetski space $W^{1/2,2}(\Gamma_i)$. Lemma 1 immediately follows from [6]. Naturally, since function $\mathbf{g}^*$ is divergence–free, it cannot be generally equal to zero everywhere on Γ_o.

Further, we construct a solution $\mathbf{u}$ of the problem (1)–(5) in the form

$$\mathbf{u} = \mathbf{g}^* + \omega\,\mathbf{x}^\perp + \mathbf{v}, \tag{7}$$

where $\mathbf{v}$ is a new unknown function. Substituting this to (1)–(5), we obtain the classical formulation of the considered boundary–value problem:

$$\big((\mathbf{g}^* + \mathbf{v})\cdot\nabla\big)\mathbf{v} + (\mathbf{v}\cdot\nabla)\mathbf{g}^* + 2\omega\,\mathbf{v}^\perp = \mathbf{h} - \nabla p + \nu\,\Delta\mathbf{v} \quad \text{in } \Omega, \tag{8}$$

$$\operatorname{div}\mathbf{v} = 0 \quad \text{in } \Omega, \tag{9}$$

$$\mathbf{v} = \mathbf{0} \quad \text{on } \Gamma_i \cup \Gamma_w, \tag{10}$$

$$-\nu\,\frac{\partial\mathbf{v}}{\partial\mathbf{n}} + p\,\mathbf{n} - \frac{1}{2}\,\big[(\mathbf{g}^* + \mathbf{v})\cdot\mathbf{n}\big]^-\,(\mathbf{g}^* + \mathbf{v}) = \mathbf{q} \quad \text{on } \Gamma_o, \tag{11}$$

where

$$\mathbf{h} = \mathbf{f} - (\mathbf{g}^*\cdot\nabla)\mathbf{g}^* - 2\omega\,\mathbf{g}^{*\perp} + \omega^2\,\mathbf{x} + \nu\,\Delta\mathbf{g}^*, \quad \mathbf{q} = \nu\,\frac{\partial\mathbf{g}^*}{\partial\mathbf{n}} - \nu\omega\,\frac{\mathbf{x}^\perp}{|\mathbf{x}|}. \tag{12}$$

The term $2\omega \mathbf{v}^\perp$ in Eq. (8) represents the Coriolis force and the term $\omega^2 \mathbf{x}$ in function $\mathbf{h}$ expresses the centrifugal force. One may logically expect that the term with the negative part in (11) should have the form $[(\mathbf{g}^* + \omega \mathbf{x}^\perp + \mathbf{v}) \cdot \mathbf{n}]^-$. However, it can be simplified to $[(\mathbf{g}^* + \mathbf{v}) \cdot \mathbf{n}]^-$ because $\omega \mathbf{x}^\perp \cdot \mathbf{n} = 0$ on Γ_o.

3 A Weak Solution

3.1 Weak Formulation of the Problem (8)–(11)

We use the function space

$$V := \{\mathbf{v} \in W^{1,2}(\Omega)^2;\ \mathrm{div}\, \mathbf{v} = 0 \text{ a.e. in } \Omega,\ \mathbf{v} = \mathbf{0} \text{ on } \Gamma_i \cup \Gamma_w\}$$

with the norm

$$|||\mathbf{v}||| := \left(\int_\Omega |\nabla \mathbf{v}|^2 \, \mathrm{d}\mathbf{x} \right)^{1/2}.$$

The norms $|||\,.\,|||$ and $\|\,.\,\|_{1,2}$ are equivalent in V. If we formally multiply Eq. (8) by an arbitrary test function $\mathbf{w} = (w_1, w_2) \in V$, integrate in Ω, apply the integration by parts and use conditions (9)–(11), we obtain

$$\int_\Omega \Big[\nu\, \nabla \mathbf{v} : \nabla \mathbf{w} + \big((\mathbf{g}^* + \mathbf{v}) \cdot \nabla\big)\mathbf{v} \cdot \mathbf{w} + (\mathbf{v} \cdot \nabla)\mathbf{g}^* \cdot \mathbf{w} + 2\omega\, \mathbf{v}^\perp \cdot \mathbf{w} \Big] \, \mathrm{d}\mathbf{x}$$
$$+ \frac{1}{2} \int_{\Gamma_o} \big[(\mathbf{g}^* + \mathbf{v}) \cdot \mathbf{n}\big]^- (\mathbf{g}^* + \omega \mathbf{x}^\perp + \mathbf{v}) \cdot \mathbf{w} \, \mathrm{d}l + \int_{\Gamma_o} \mathbf{q} \cdot \mathbf{w} \, \mathrm{d}l \ = \ \int_\Omega \mathbf{h} \cdot \mathbf{w} \, \mathrm{d}\mathbf{x}. \tag{13}$$

In order to write Eq. (13) in a simple form, we define

$$a_1(\mathbf{u}, \mathbf{v}) := \nu \int_\Omega \nabla \mathbf{u} : \nabla \mathbf{v} \, \mathrm{d}\mathbf{x},$$

$$a_2(\mathbf{u}, \mathbf{v}, \mathbf{w}) := \int_\Omega (\mathbf{u} \cdot \nabla)\mathbf{v} \cdot \mathbf{w} \, \mathrm{d}\mathbf{x},$$

$$a_3(\mathbf{u}, \mathbf{v}, \mathbf{w}) := \int_{\Gamma_o} \frac{1}{2} (\mathbf{u} \cdot \mathbf{n})^- \mathbf{v} \cdot \mathbf{w} \, \mathrm{d}l,$$

$$a(\mathbf{v}, \mathbf{w}) := a_1(\mathbf{v}, \mathbf{w}) + a_2(\mathbf{g}^* + \mathbf{v}, \mathbf{v}, \mathbf{w}) + a_2(\mathbf{v}, \mathbf{g}^*, \mathbf{w})$$
$$+ 2\omega\, (\mathbf{v}^\perp, \mathbf{w})_{2;\Omega} + a_3(\mathbf{g}^* + \mathbf{v}, \mathbf{g}^* + \mathbf{v}, \mathbf{w}),$$

$$b(\mathbf{w}) := -(\mathbf{q}, \mathbf{w})_{2;\Gamma_o} + (\mathbf{f}, \mathbf{w})_{2;\Omega},$$

where $(.,.)_{2;\Omega}$ (respectively $(.,.)_{2;\Gamma_o}$) is the scalar product in $L^2(\Omega)^2$ (respectively in $L^2(\Gamma_o)^2$). Using this notation, we can write the integral equation (13) in the form

$$a(\mathbf{v},\mathbf{w}) \;=\; b(\mathbf{w}). \tag{14}$$

A function $\mathbf{v} \in V$ is said to be a *weak solution* of the problem (8)–(11) if it satisfies Eq. (14) for all $\mathbf{w} \in V$. The pressure p does not explicitly appear in the definition of the weak solution. However, as it is usual in the theory of the Navier–Stokes equations, it can be defined on the level of distributions.

3.2 Auxiliary Estimates

The next theorem plays a fundamental role in the proof of the existence of a weak solution.

Theorem 1. *There exist positive constants c_2 and c_3, depending on domain Ω and the angular velocity ω, such that*

$$a(\mathbf{v},\mathbf{v}) \;\geq\; |||\mathbf{v}|||\left(\nu\,|||\mathbf{v}||| - c_2\,\|\mathbf{g}\|^2_{1/2;\,\Gamma_i} - c_3\,\|\mathbf{g}\|_{1/2;\,\Gamma_i}\,|||\mathbf{v}|||\right) \tag{15}$$

for all $\mathbf{v} \in V$.

Proof. Using the definition of the forms a_1, a_2 and a_3 in Sect. 3.1, we derive the estimates:

$$a_1(\mathbf{v},\mathbf{v}) \;:=\; \nu\int_\Omega \nabla\mathbf{v} : \nabla\mathbf{v}\,\mathrm{d}\mathbf{x} \;=\; \nu\,|||\mathbf{v}|||^2, \tag{16}$$

$$|a_2(\mathbf{g}^*,\mathbf{v},\mathbf{v})| \;=\; \left|\int_\Omega (\mathbf{g}^*\cdot\nabla)\mathbf{v}\cdot\mathbf{v}\,\mathrm{d}\mathbf{x}\right| \;\leq\; |||\mathbf{v}|||\left(\int_\Omega |\mathbf{g}^*|^2\,|\mathbf{v}|^2\,\mathrm{d}\mathbf{x}\right)^{1/2}$$

$$\leq\; C\,|||\mathbf{v}|||\,\|\mathbf{g}^*\|_4\,\|\mathbf{v}\|_4 \;\leq\; C\,|||\mathbf{v}|||^2\,\|\mathbf{g}^*\|_{1,2}, \tag{17}$$

$$|a_2(\mathbf{v},\mathbf{g}^*,\mathbf{v})| \;=\; \left|\int_\Omega (\mathbf{v}\cdot\nabla)\mathbf{g}^*\cdot\mathbf{v}\,\mathrm{d}\mathbf{x}\right| \;\leq\; \|\mathbf{v}\|^2_4\,\|\mathbf{g}^*\|_{1,2} \;\leq\; C\,|||\mathbf{v}|||^2\,\|\mathbf{g}^*\|_{1,2}, \tag{18}$$

$$a_2(\mathbf{v},\mathbf{v},\mathbf{v}) + a_3(\mathbf{g}^*+\mathbf{v},\mathbf{v},\mathbf{v}) = \int_\Omega \big((\mathbf{g}^*+\mathbf{v})\cdot\nabla\big)\mathbf{v}\cdot\mathbf{v}\,\mathrm{d}\mathbf{x}$$

$$-\int_\Omega (\mathbf{g}^*\cdot\nabla)\mathbf{v}\cdot\mathbf{v}\,\mathrm{d}\mathbf{x} \;+\; \frac{1}{2}\int_{\Gamma_o}\big((\mathbf{g}^*+\mathbf{v})\cdot\mathbf{n}\big)^-\,|\mathbf{v}|^2\,\mathrm{d}l$$

$$=\; \frac{1}{2}\int_\Omega \big((\mathbf{g}^*+\mathbf{v})\cdot\nabla\big)|\mathbf{v}|^2\,\mathrm{d}\mathbf{x} - \frac{1}{2}\int_\Omega (\mathbf{g}^*\cdot\nabla)|\mathbf{v}|^2\,\mathrm{d}\mathbf{x}$$

$$
\begin{aligned}
&+\frac{1}{2}\int_{\Gamma_o}\big((\mathbf{g}^*+\mathbf{v})\cdot\mathbf{n}\big)^-|\mathbf{v}|^2\,\mathrm{d}l\\
&=\frac{1}{2}\int_{\Gamma_o}\big((\mathbf{g}^*+\mathbf{v})\cdot\mathbf{n}\big)|\mathbf{v}|^2\,\mathrm{d}l-\frac{1}{2}\int_{\Gamma_o}(\mathbf{g}^*\cdot\mathbf{n})|\mathbf{v}|^2\,\mathrm{d}l\\
&\quad-\frac{1}{2}\int_{\Omega}\mathrm{div}\,(\mathbf{g}^*+\mathbf{v})\,|\mathbf{v}|^2\,\mathrm{d}\mathbf{x}+\frac{1}{2}\int_{\Omega}\mathrm{div}\,\mathbf{g}^*\,|\mathbf{v}|^2\,\mathrm{d}\mathbf{x}\\
&\quad+\frac{1}{2}\int_{\Gamma_o}\big((\mathbf{g}^*+\mathbf{v})\cdot\mathbf{n}\big)^-|\mathbf{v}|^2\,\mathrm{d}l\\
&=\frac{1}{2}\int_{\Gamma_o}\big((\mathbf{g}^*+\mathbf{v})\cdot\mathbf{n}\big)|\mathbf{v}|^2\,\mathrm{d}l-\frac{1}{2}\int_{\Gamma_o}(\mathbf{g}^*\cdot\mathbf{n})|\mathbf{v}|^2\,\mathrm{d}l\\
&\quad+\frac{1}{2}\int_{\Gamma_o}\big((\mathbf{g}^*+\mathbf{v})\cdot\mathbf{n}\big)^-|\mathbf{v}|^2\,\mathrm{d}l\\
&=\frac{1}{2}\int_{\Gamma_o}\big((\mathbf{g}^*+\mathbf{v})\cdot\mathbf{n}\big)^+|\mathbf{v}|^2\,\mathrm{d}l-\frac{1}{2}\int_{\Gamma_o}(\mathbf{g}^*\cdot\mathbf{n})|\mathbf{v}|^2\,\mathrm{d}l\\
&\geq-\frac{1}{2}\int_{\Gamma_o}(\mathbf{g}^*\cdot\mathbf{n})|\mathbf{v}|^2\,\mathrm{d}\ \geq-\|\mathbf{g}^*\|_{2;\,\Gamma_o}\,\|\mathbf{v}\|^2_{4;\,\Gamma_o}\ \geq-\|\mathbf{g}^*\|_{1,2}\,|||\mathbf{v}|||^2, \qquad (19)
\end{aligned}
$$

$$
\begin{aligned}
|a_3(\mathbf{g}^*+\mathbf{v},\mathbf{g}^*,\mathbf{v})|\ &=\ \left|\frac{1}{2}\int_{\Gamma_o}\big((\mathbf{g}^*+\mathbf{v})\cdot\mathbf{n}\big)^-\mathbf{g}^*\cdot\mathbf{v}\,\mathrm{d}l\right|\\
&\leq\frac{1}{2}\int_{\Gamma_o}|\mathbf{g}^*|^2\,|\mathbf{v}|\,\mathrm{d}l+\frac{1}{2}\int_{\Gamma_o}|\mathbf{g}^*|\,|\mathbf{v}|^2\,\mathrm{d}l\\
&\leq C\,\|\mathbf{g}^*\|^2_{4;\,\Gamma_o}\,\|\mathbf{v}\|_{2;\,\Gamma_o}+C\,\|\mathbf{g}^*\|_{2;\,\Gamma_o}\,\|\mathbf{v}\|^2_{4;\,\Gamma_o}\\
&\leq C\,\|\mathbf{g}^*\|^2_{1,2}\,|||\mathbf{v}|||+C\,\|\mathbf{g}^*\|_{1,2}\,|||\mathbf{v}|||^2, \qquad (20)
\end{aligned}
$$

$$
2\omega\,(\mathbf{v}^\perp,\mathbf{v})_{2;\,\Omega}\ =\ 0. \qquad (21)
$$

We denote by C the generic constant, i.e. the constant which may change the value from line to line. It depends on Ω and ω, but it is independent of $\mathbf{g}^*$ and $\mathbf{v}$. Substituting now from inequalities (16)–(21) to the definition of the integral form a (see Sect. 3.1), and estimating the norm $\|\mathbf{g}^*\|_{1,2}$ by means of (6), we obtain inequality (15).

Theorem 2. *If function* $\mathbf{g}$*, defined on* Γ_i*, is so small that*

$$
c_3\,\|\mathbf{g}\|_{1/2;\,\Gamma_i}\ <\ \nu \qquad (22)
$$

(where c_3 *is the constant from Theorem 1) then the form* $a(\mathbf{v},\mathbf{v})$ *is coercive in space* V*. It means that there exist positive constants* c_4*,* c_5 *such that*

$$a(\mathbf{v},\mathbf{v}) \;\geq\; c_4 \,|||\mathbf{v}|||^2 \tag{23}$$

for $\mathbf{v} \in V$ *such that* $|||\mathbf{v}||| \geq c_5$.

Proof. If we denote $c_6 := \nu - c_3 \,\|\mathbf{g}\|_{1/2;\,\Gamma_i}$ then estimate (15) yields

$$a(\mathbf{v},\mathbf{v}) = \geq \; c_6 \,|||\mathbf{v}|||^2 - c_2 \,\|\mathbf{g}\|^2_{1/2;\,\Gamma_i} \,|||\mathbf{v}|||. \tag{24}$$

The validity of (23) for sufficiently large $\mathbf{v}$ now follows from (24).

Theorem 3 (on the existence of a weak solution). *Let* $\omega \in \mathbb{R}$. *Suppose that the norm* $\|\mathbf{g}\|_{1/2;\,\Gamma_i}$ *is so small that it satisfies inequality (22). Then there exists a weak solution to the problem (8)–(11).*

The proof of this theorem is based on the construction, estimates and convergence of appropriate Galerkin's approximation (see for e.g. [7]).

Conclusion. Recall that $\mathbf{v}$ is the weak solution of the problem, formulated in the rotating frame, and the considered domain in the rotating frame is time–independent. The function $\mathbf{u} := \mathbf{g}^* + \omega\mathbf{x}^\perp + \mathbf{v}$ (see (7)), is a weak solution of the problem (1)–(5), which is also formulated in the rotating frame. The function $R^{-1}(t)\;\mathbf{u}\big(R(t)\mathbf{x}\big)$ is a solution of the problem, described in a fixed frame. (The domain, occupied by the moving fluid, is naturally time dependent from the point of view of an observer, whose position is fixed outside the rotating turbine.)

Acknowledgements The research was supported by the research plan of the Ministry of Education of the Czech Republic No. MSM 6840770010 and by the Grant Agency of the Czech Rep., grant No. 201/09/P413.

References

1. Bruneau C. H., Fabrie P.: New efficient boundary conditions for incompressible Navier–Stokes equations: A well–posedness result. *Mathematical Modelling and Numerical Analysis* 30, No. 7, 815–840, 1996.
2. Hishida T.: An existence theorem for the Navier–Stokes flow in the exterior of a rotating obstacle. *Arch. Rational Mech. Anal.* 150, 307–348, 1999.
3. Farwig R., Neustupa J.: On the spectrum of a Stokes–type operator arising from flow around a rotating body. *Manuscripta Mathematica* 122, 419–437, 2007.
4. Glowinski R.: *Numerical Methods for Nonlinear Variational Problems.* Springer–Verlag, New York-Berlin-Heidelberg-Tokyo, 1984.
5. Kučera P., Skalák Z.: Local solutions to the Navier-Stokes equations with mixed boundary conditions. *Acta Applicandae Mathematicae* 54, No. 3, 275–288, 1998.
6. Feistauer M., Neustupa T.: On some aspects of analysis of incompressible flow through cascades of profiles. *Operator Theory, Advances and Applications,* Vol. 147, Birkhauser, Basel, 257–276, 2004.
7. Temam R.: *Navier-Stokes Equations.* North-Holland, Amsterdam-New York-Oxford, 1977.

Combined Mixed-Hybrid Finite Element–Finite Volume Scheme for Computation of Multicomponent Compressible Flow in Porous Media

O. Polívka and J. Mikyška

Abstract The paper deals with the numerical modeling of compressible single-phase flow of a mixture composed of several components in a porous medium. The mathematical model is formulated by Darcy's law, components continuity equations, constitutive relations, and initial and boundary conditions. The problem is solved numerically using a combination of the mixed-hybrid finite element method for the total flux discretization and the finite volume method for the discretization of the transport equations. The time discretization is carried out by Euler's method. The resulting large system of nonlinear algebraic equations is solved by the Newton-Raphson method. The dimensions of the obtained system of linear algebraic equations are significantly reduced so that they do not depend on the number of mixture components. The convergence of the numerical scheme is verified in the single-component case by comparing the numerical solution with an analytical solution.

1 Introduction

The mathematical modeling of the transport of multicomponent mixtures in the subsurface is important for many applications including oil recovery or CO_2 sequestration. The traditional approaches use either the fully implicit (fully coupled) method or a sequential method [5,14]. The fully implicit method is stable, allows for long time steps, but leads to extremely large systems of linear algebraic equations whose size is proportional to the number of mixture components. Alternatively, in sequential solution procedures like IMPEC (implicit pressure, explicit concentrations) [8], a pressure equation is formulated by summing up the transport equations

O. Polívka (✉) · J. Mikyška
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: ondrej.polivka@fjfi.cvut.cz; jiri.mikyska@fjfi.cvut.cz

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_59,

[5, 14] or by another method [1, 7, 15]. This procedure allows the size of the solved system to be reduced, as only pressure is solved implicitly. However, this approach is conditionally stable and the time step has to be chosen prohibitively small in many cases.

In this paper, we improve our approach to the numerical modeling of the compressible multicomponent single-phase flow in a porous medium proposed in [13], where the numerical scheme was used for a simulation of methane injection into a propane reservoir. The original approach handled a velocity discretization; now, the total flux is discretized, and the convergence of the numerical to an analytical solution in a special case is verified. The scheme, based on a combination of the mixed-hybrid finite element method (MHFEM) and the finite volume method (FVM), has advantages of both the traditional sequential and implicit methods. As in the implicit schemes, our method leads to large systems of linear algebraic equations, but it is possible to reduce the size of the final system of equations to a size independent of the number of mixture components. Unlike in other sequential approaches, no pressure equation has to be formed as pressure is evaluated directly from the equation of state.

2 Mathematical Model

Let $\Omega \subset \mathbb{R}^2$ be a bounded domain with porosity ϕ [-], and (t_0, τ) be the time interval [s]. Consider the single-phase compressible flow of a fluid of n_c components in the domain at a constant temperature T [K]. Neglecting diffusion, the transport of the components is described by the following molar balance equations [7]

$$\frac{\partial(\phi c_i)}{\partial t} + \nabla \cdot (c_i \mathbf{v}) = f_i, \qquad i = 1, \ldots, n_c, \tag{1}$$

$$c_i = c_i(\mathbf{x}, t), \qquad \mathbf{x} \in \Omega,\ t \in (t_0, \tau),$$

$$\mathbf{q} = c\mathbf{v}, \qquad c = \sum_{i=1}^{n_c} c_i, \tag{2}$$

where unknown quantities c_i, $i = 1, \ldots, n_c$, are the molar concentrations of the components [mol m^{-3}]. On the right hand side of Eq. (1), f_i [mol m^{-3} s^{-1}] denotes the sink/source term. The total molar flux $\mathbf{q}$ is expressed in (2) by the total molar concentration c and Darcy's velocity $\mathbf{v}$ [m s^{-1}] which is given according to [2] by

$$\mathbf{v} = -\mu^{-1}\mathsf{K}(\nabla p - \varrho \mathbf{g}). \tag{3}$$

In (3), K is the medium intrinsic permeability [m^2] (generally symmetric and uniformly positive-definite tensor), μ is the viscosity [kg m^{-1} s^{-1}], ∇p denotes a gradient of the pressure p [Pa], $\mathbf{g}$ is the gravitational acceleration vector [m s^{-2}], and

ϱ is the fluid density [$\mathrm{kg\,m^{-3}}$]. Equations (1) and (3) are coupled with constitutive relations expressing dependencies (to be found in [6, 10, 12, 13])

$$p = p\left(c_1, \ldots, c_{n_c}, T\right), \quad \varrho = \varrho\left(c_1, \ldots, c_{n_c}\right), \quad \mu = \mu\left(c_1, \ldots, c_{n_c}, T\right). \tag{4}$$

The initial and boundary conditions are given by

$$c_i(\mathbf{x}, t_0) = c_i^0(\mathbf{x}), \qquad \mathbf{x} \in \Omega,\ i = 1, \ldots, n_c, \tag{5a}$$

$$c_i(\mathbf{x}, t) = c_i^D(\mathbf{x}, t), \qquad \mathbf{x} \in \Gamma_c(t),\ t \in (t_0, \tau),\ i = 1, \ldots, n_c, \tag{5b}$$

$$p(\mathbf{x}, t) = p^D(\mathbf{x}, t), \qquad \mathbf{x} \in \Gamma_p,\ t \in (t_0, \tau), \tag{5c}$$

$$\mathbf{q}(\mathbf{x}, t) \cdot \mathbf{n}(\mathbf{x}) = q^N(\mathbf{x}, t), \qquad \mathbf{x} \in \Gamma_q,\ t \in (t_0, \tau), \tag{5d}$$

where $\mathbf{n}$ is the unit outward normal vector to the boundary $\partial\Omega$, $\Gamma_p \cup \Gamma_q = \partial\Omega$, and $\Gamma_p \cap \Gamma_q = \emptyset$. Further, $\Gamma_c(t)$ denotes the inflow part of the boundary $\partial\Omega$ at time t, i.e. $\Gamma_c(t) = \left\{\mathbf{x} \in \partial\Omega \mid \mathbf{q}(\mathbf{x}, t) \cdot \mathbf{n}(\mathbf{x}) < 0\right\}$. On $\Gamma_c \cap \Gamma_p$, values of c_i^D, $i = 1, \ldots, n_c$, are constrained by Eqs. (4) and (5c) so that $p^D = p\left(c_1^D, \ldots, c_{n_c}^D, T\right)$.

3 Numerical Scheme

The system of Eqs. (1)–(5) is solved numerically by a combination of the MHFEM, for total flux relation (2), and the FVM, for transport Eqs. (1). We consider a 2D polygonal domain Ω with the boundary $\partial\Omega$ which is covered by a conforming triangulation T_Ω. Let us denote K the element of the mesh T_Ω with area $|K|$, E the edge of an element with the length $|E|$, n_k the number of elements of the triangulation, and n_e the number of edges of the mesh.

Discretization of the Total Molar Flux Unlike in [13], where the velocity $\mathbf{v}$ was discretized, here, the total molar flux $\mathbf{q}$ is approximated in the Raviart-Thomas space of the lowest order (RT_K^0) over the element $K \in T_\Omega$ as

$$\mathbf{q} = \sum_{E \in \partial K} q_{K,E} \mathbf{w}_{K,E}, \tag{6}$$

where the coefficient $q_{K,E}$ is the numerical flux of vector function $\mathbf{q}$ through the edge E of the element K with respect to outer normal, and $\mathbf{w}_{K,E}$ represents the piecewise linear RT_K^0-basis function associated with the edge E (see [3, 4, 11, 13]).

If we express the pressure gradient from Darcy's law (3), multiply both sides of the obtained relation by the basis function $\mathbf{w}_{K,E}$, integrate over K, use (6) and properties of the RT_K^0 space, we derive a discrete form of (2) and (3)

$$q_{K,E} = c_K \mu_K^{-1} \left(\alpha_{K,E} p_K - \sum_{E' \in \partial K} \beta_{K,E,E'} p_{K,E'} + \gamma_{K,E} \varrho_K \right), \quad E \in \partial K. \tag{7}$$

Green's theorem and the mean value theorem were also employed in the derivation [13]. In (7), $\alpha_{K,E}, \beta_{K,E,E'}$, and $\gamma_{K,E}$ are coefficients dependent on the mesh geometry and on the local values of permeability (details in [13]); p_K is the cell pressure average, $p_{K,E'}$ is the edge pressure average, and μ_K, ϱ_K denote the mean values of viscosity and density over the cell K, respectively.

The continuity of the flux and pressure on the edge E between neighboring elements $K, K' \in T_\Omega$ can be written as

$$q_{K,E} + q_{K',E} = 0, \quad p_{K,E} = p_{K',E} =: p_E. \tag{8}$$

Boundary conditions (5c) and (5d) in a discrete form read as

$$p_{K,E} = p^D(E), \quad \forall E \subset \Gamma_p, \tag{9a}$$

$$q_{K,E} = q^N(E), \quad \forall E \subset \Gamma_q, \tag{9b}$$

where $p^D(E)$ is the prescribed value of the pressure p averaged on the edge E, and $q^N(E)$ is prescribed flux through the edge E.

The numerical fluxes can be eliminated by substituting $q_{K,E}$ from (7) into (8) and (9b). For further derivation, let us consider time dependent quantities at time t_{n+1} denoted by upper index $n+1$. Then, Eqs. (7)–(9) transform to the following system of n_e linear algebraic equations

$$F_E \equiv \begin{cases} \displaystyle\sum_{K:E\in\partial K} c_K^{n+1} \left(\mu_K^{n+1}\right)^{-1} \left(\alpha_{K,E} p_K^{n+1} - \sum_{E'\in\partial K} \beta_{K,E,E'} p_{K,E'}^{n+1} + \gamma_{K,E} \varrho_K^{n+1} \right) \\ \qquad\qquad - \displaystyle\sum_{K:E\in\partial K\cap\Gamma_q} q^N(E) = 0, \quad \forall E \not\subset \Gamma_p, \\ p_{K,E}^{n+1} - p^D(E) = 0, \quad \forall E \subset \Gamma_p \,. \end{cases} \tag{10}$$

Herein, $\sum_{K:E\in\partial K}$ denotes the sum over the elements adjacent to the edge E.

Approximation of the Transport Equations Transport Eqs. (1) with the initial and boundary conditions (5) are discretized by the FVM [9]. Equation (1) is then integrated over an arbitrary element K. Using Green's theorem, applying the mean value theorem, and denoting $\phi_K, c_{i,K}, f_{i,K}$, the averaged values of ϕ, c_i, f_i ($i = 1, \ldots, n_c$) over the cell K, respectively, the discrete form of (1) reads as

$$\frac{\mathrm{d}(\phi_K c_{i,K})}{\mathrm{d}t}|K| + \sum_{E\in\partial K} \widetilde{z_{i,E}} \int_E \mathbf{q}\cdot\mathbf{n}_{K,E} = f_{i,K}|K|, \tag{11}$$

where $\widetilde{z_{i,E}}$ denotes the mole fraction $z_i = c_i/c$ of the i-th component on the edge E, and $\mathbf{q}$ is given by (2). The integral in (11) is equal to the numerical flux $q_{K,E}$.

Let us suppose that the porosity does not depend on time. The time derivative of $c_{i,K}$ in (11) is approximated by the time difference with a time step Δt_n. Using Euler's method [9], we obtain for every n, all $K \in T_\Omega$, and $i = 1, \ldots, n_c$

$$F_{K,i} \equiv \phi_K |K| \frac{c_{i,K}^{n+1} - c_{i,K}^n}{\Delta t_n} + \sum_{E \in \partial K} \widetilde{z_{i,E}}^n \, q_{K,E}^{n+1} \left(p_{K,E}^{n+1}, c_{1,K}^{n+1}, \ldots, c_{n_c,K}^{n+1} \right) - f_{i,K} |K| = 0, \tag{12}$$

where $q_{K,E}^{n+1}$ is given by (7). The value of $\widetilde{z_{i,E}}^n$ is chosen by upwinding as

$$\widetilde{z_{i,E}}^n = \begin{cases} z_{i,K}^n & \text{for} \quad q_{K,E}^{n+1} \geq 0, \\ z_{i,K'}^n & \text{for} \quad q_{K,E}^{n+1} < 0 \ \wedge \ E \not\subset \partial\Omega : K \cap K' = E, \\ z_{i,E}^{D,n} & \text{for} \quad q_{K,E}^{n+1} < 0 \ \wedge \ E \subset \partial\Omega, \end{cases} \tag{13}$$

where z_i^D represents the mole fraction of the i-th component on the inflow boundary computed from (5b). Note that the scheme is almost fully implicit, the only term in (12) which is evaluated explicitly is the value of $\widetilde{z_{i,E}}^n$.

The initial and boundary conditions (5a) and (5b) are approximated as

$$c_{i,K}^0 = c_i^0(K), \qquad \forall K \in T_\Omega, \ i = 1, \ldots, n_c, \tag{14a}$$

$$\widetilde{z_{i,E}}^n = z_i^D(E, t_n), \qquad \forall E \subset \Gamma_c(t), \ i = 1, \ldots, n_c, \ t_0 < t_n < \tau. \tag{14b}$$

Combining the MHFEM and FVM Schemes Let us denote F_E and $F_{K,i}$, for edge $E \in \{1, \ldots, n_e\}$, element $K \in \{1, \ldots, n_k\}$, and component $i \in \{1, \ldots, n_c\}$, the left hand sides of Eqs. (10) and (12) with $q_{K,E}^{n+1}$ substituted from relation (7). The cell-averaged values $p_K = p\left(c_{1,K}, \ldots, c_{n_c,K}\right)$, $\varrho = \varrho_K\left(c_{1,K}, \ldots, c_{n_c,K}\right)$, and $\mu_K = \mu\left(c_{1,K}, \ldots, c_{n_c,K}\right)$ are evaluated using (4). The system of $n_e + n_k \times n_c$ equations

$$\mathbf{F} = \left[F_1, \ldots, F_{n_e}; F_{1,1}, \ldots, F_{1,n_c}, \ldots, F_{n_k,1}, \ldots, F_{n_k,n_c} \right]^T = \mathbf{0}$$

for unknown molar concentrations $c_{1,K}^{n+1}, \ldots, c_{n_c,K}^{n+1}$, $K \in \{1, \ldots, n_k\}$, and edge-averaged pressures p_E^{n+1}, $E \in \{1, \ldots, n_e\}$, is a nonlinear system of algebraic equations which we solve using the Newton-Raphson method (NRM). The resulting system of linear algebraic equations is shown in Fig. 1, where the sparse Jacobi matrix is unsymmetric, and the unknown vector is represented by corrections of molar concentrations and edge pressures. The nonzero black-colored blocks in Fig. 1 are given by partial derivatives

$$(\mathbf{J}_K)_{i,j} = \frac{\partial F_{K,i}}{\partial c_{j,K}^{n+1}}, \ \left(\mathbf{J}_{K,E}\right)_i = \frac{\partial F_{K,i}}{\partial p_{K,E}^{n+1}}, \ \left(\mathbf{J}_{E,K}\right)_j = \frac{\partial F_E}{\partial c_{j,K}^{n+1}}, \ J_{E,E'} = \frac{\partial F_E}{\partial p_{K,E'}^{n+1}}, \tag{15}$$

where $J_{E,E'}$ is element of $\mathbf{J}_{E,E'}$ and $i, j = 1, \ldots, n_c$; $K = 1, \ldots, n_k$; $E, E' = 1, \ldots, n_e$. The partial derivatives in (15) can be evaluated analytically using (4), (10), and (12).

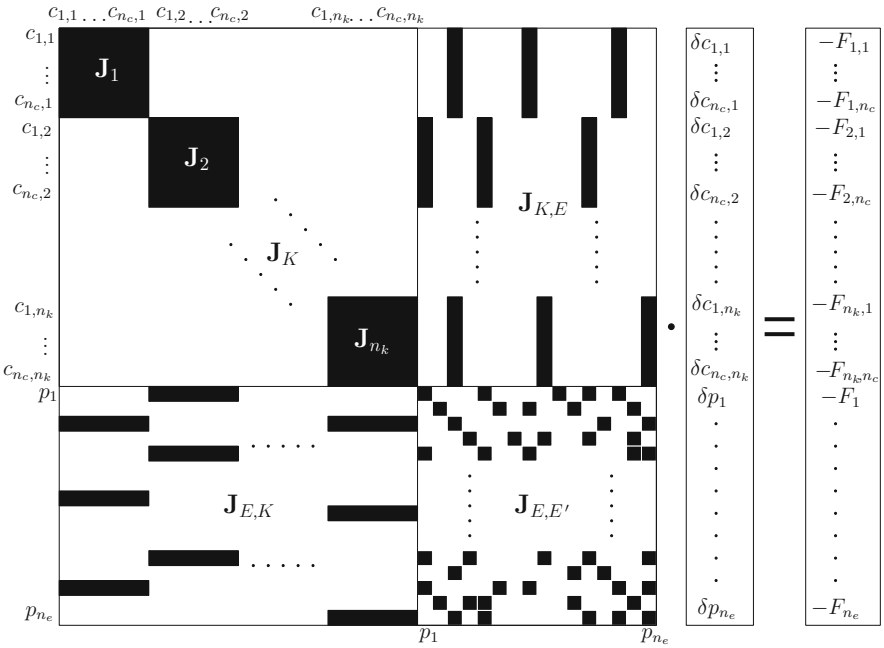


Fig. 1 Structure of the system of linear algebraic equations in the NRM

The size of the system in Fig. 1 can be reduced by inverting the $\mathbf{J}_K$ blocks for all K (the inversion is possible since the blocks are diagonally dominant for small time steps) and eliminating vectors $\mathbf{J}_{E,K}$ for all E, K. Thus, we derive a reduced system of n_e equations for n_e corrections of pressures δp_E with the same structure as $\mathbf{J}_{E,E'}$. Once δp_E are computed, corrections of concentrations $\delta c_{1,K}, \ldots, \delta c_{n_c,K}$ on each cell K can be evaluated by the back-substitution utilizing the $\mathbf{J}_K$ inversions.

4 Experimental Analysis of Convergence

In this section, we verify convergence of the proposed scheme using the experimental convergence analysis. Choosing $n_c = 1, \phi = 1, f = 0, \mathbf{g} = 0, \mathsf{K} = 1, \mu = 0.5$, and $p = c$, Eqs. (1) and (3) transform to

$$\frac{\partial c}{\partial t} + \nabla \cdot (c\mathbf{v}) = 0, \quad \mathbf{v} = -2\,\nabla c. \tag{16}$$

We set the initial and boundary conditions as follows

$$c(\mathbf{x}, t_0) = B_2(x, t_0), \qquad \mathbf{x} \in \Omega, \tag{17a}$$

$$p(\mathbf{x}, t) = B_2(x, t), \qquad \mathbf{x} \in \Gamma_p,\ t \in (t_0, \tau)\,, \tag{17b}$$

$$\mathbf{q}(\mathbf{x}, t) \cdot \mathbf{n}(\mathbf{x}) = 0, \qquad \mathbf{x} \in \Gamma_q,\ t \in (t_0, \tau)\,, \tag{17c}$$

where $B_2(x, t)$ is the Barenblatt solution of (16) prescribed by

$$B_m(x, t) = t^{-k} \left[\left(1 - \frac{k(m-1)}{2m} \frac{|x|^2}{t^{2k}} \right)_+ \right]^{1/(m-1)}, \quad m = 2, \tag{18}$$

where $c_+ = \max(c, 0)$ and $k = (m+1)^{-1}$. For any time $t > 0$, solution (18) has a compact support $[-t^k \sqrt{2m/(k(m-1))},\ t^k \sqrt{2m/(k(m-1))}\,]$ becoming wider in a finite speed [16].

Using the numerical scheme derived in Sect. 3, we solve Eqs. (16) and (17) in a rectangular domain $100 \times 20\,\mathrm{m}^2$, where Γ_p is composed of line segments $x = 0$ and $x = 100$ ($0 \le y \le 20$), while Γ_q contains the rest of the boundary $\partial\Omega$, i.e. the horizontal and vertical boundary $y = 0$ and $y = 20$. Both parts of Γ_p are outflow boundaries, thus $\Gamma_c = \emptyset$. The initial time $t_0 = 10^4$ s and the final time $\tau = 10^6$ s.

The numerical solution is compared with the analytical one (18) by means of the experimental orders of convergence (EOCs). Every numerical (element-wise constant) solution is computed on a homogeneous regular triangular grid containing $n = 2 \times n_x \times n_y$ triangles, and projected to a grid on which the analytical solution is computed. The analytical solution is element-wise linear by evaluating from (18) three values on edges of each element. Error E_n between the analytical solution and the projection of the numerical solution (to the grid of the analytical solution) originally computed on the grid n in three consistent norms L^1, L^2, and L^∞ is evaluated.

In Table 1, EOCs and errors of concentration on five grids with $n_x = n_y$ and the finest $n_x = 320$ are included. The analytical solution is interpolated on the grid $n_x = 640$. The time step for the solution $n_x = 320$ is chosen constant $\Delta t = 386.7$ s. On every coarser grid, Δt is four times larger with each mesh refinement (i.e. $\Delta t \sim 1/n$) to observe EOC of the space discretization.[1]

The numerical solutions computed on the grids with a different n_x and n_y are compared with the analytical solution interpolated on the grid of $2 \times 1{,}600 \times 320$ cells in Table 2. Again, Δt is four times larger with each mesh refinement (i.e. $\Delta t \sim 1/n$)[1], $\Delta t = 386.7$ s for the numerical solution with $n = 2 \times 800 \times 160$.

[1]For $\Delta t \sim 1/\sqrt{n}$ all EOCs equal to 1 were observed in L^1 and L^2 norms, and the error of the time discretization, thus, prevailed.

Table 1 EOCs and errors of concentration c at time $\tau = 10^6$ s compared with the analytical solution on the grid $n_x = 640$ ($n = 2 \times n_x \times n_x$ elements) and the time step $\Delta t = 386.7$ s for the numerical solution $n_x = 320$. On coarser grids, $\Delta t \sim 1/n$

Grid (n)	$\|E_n\|_1$	EOC_1	$\|E_n\|_2$	EOC_2	$\|E_n\|_\infty$	EOC_∞
$2 \times 20 \times 20$	2.5113×10^{-2}		6.1743×10^{-4}		2.8684×10^{-5}	
		1.8297		1.7739		1.1119
$2 \times 40 \times 40$	7.0646×10^{-3}		1.8055×10^{-4}		1.3271×10^{-5}	
		1.4607		1.3506		1.0817
$2 \times 80 \times 80$	2.5667×10^{-3}		7.0798×10^{-5}		6.2704×10^{-6}	
		1.2083		1.1165		1.1570
$2 \times 160 \times 160$	1.1108×10^{-3}		3.2653×10^{-5}		2.8119×10^{-6}	
		1.1040		1.0319		1.3770
$2 \times 320 \times 320$	5.1678×10^{-4}		1.5969×10^{-5}		1.0827×10^{-6}	

Table 2 EOCs and errors of concentration c at time $\tau = 10^6$ s compared with the analytical solution on the grid $n = 2 \times 1{,}600 \times 320$ and the time step $\Delta t = 386.7$ s for the numerical solution $n = 2 \times 800 \times 160$. On coarser grids, $\Delta t \sim 1/n$

Grid (n)	$\|E_n\|_1$	EOC_1	$\|E_n\|_2$	EOC_2	$\|E_n\|_\infty$	EOC_∞
$2 \times 50 \times 10$	2.3185×10^{-2}		5.675×10^{-4}		1.8943×10^{-5}	
		2.0520		2.0493		1.7276
$2 \times 100 \times 20$	5.5911×10^{-3}		1.3711×10^{-4}		5.7199×10^{-6}	
		1.8051		1.7690		1.1810
$2 \times 200 \times 40$	1.5999×10^{-3}		4.0231×10^{-5}		2.5228×10^{-6}	
		1.5337		1.4358		1.1608
$2 \times 400 \times 80$	5.5262×10^{-4}		1.4871×10^{-5}		1.1283×10^{-6}	
		1.2957		1.1655		1.3789
$2 \times 800 \times 160$	2.2510×10^{-4}		6.6297×10^{-6}		4.3386×10^{-7}	

5 Conclusion

In this work, we have developed a numerical scheme based on a combination of the MHFEM and FVM for simulation of single-phase compressible multicomponent flow in a porous medium. We proposed a technique reducing significantly the system into a size that is independent of the number of mixture components. Consequently, computational costs are comparable with the traditional sequential approaches. Our method provides an exact local mass balance (up to the non-linear solver error) which is important for solving problems especially in a heterogeneous medium. Convergence of the numerical scheme was verified by evaluating EOCs in a special case of the problem for which an analytical solution is known. The EOCs range from 1.03 to 2.05 behaving similarly in both examined cases.

Acknowledgements This research has been supported by the projects: Development of Computational Models for Simulation of CO2 Sequestration, P105/11/1507 of the Czech Science Foundation, Numerical Methods for Multiphase Flow and Transport in Subsurface Environmental Applications, Kontakt ME10009, and Applied Mathematics in Technical and Physical Sciences, Research Direction Project No. MSM6840770010, both of the Czech Ministry of Education.

References

1. Acs, G., Doleschall, S., Farkas, E.: General Purpose Compositional Model, *Society of Petroleum Engineers Journal*, Vol.: 25, Issue: 4 (1985) 543–553.
2. Bear, J., Verruijt, A.: Modeling Groundwater Flow and Pollution (1987), D. Reidel Publishing Company, Dordrecht, Holland.
3. Brezzi, F., Fortin, M.: Mixed and Hybrid Finite Element Methods (1991), Springer-Verlag, New York Inc.
4. Chavent, G., Roberts, J. E.: A unified physical presentation of mixed, mixed-hybrid finite elements and standard finite difference approximations for the determination of velocities in waterflow problems, *Advances in Water Resources*, 14(6) (1991).
5. Chen, Z., Huan, G., Ma, Y.: Computational Methods for Multiphase Flows in Porous Media (2006), SIAM, Philadelphia.
6. Holzbecher, E. O.: Modeling Density-Driven Flow in Porous Media: Principles, Numerics, Software (1998), Springer-Verlag, Berlin.
7. Hoteit, H., Firoozabadi, A.: Multicomponent Fluid Flow by Discontinuous Galerkin and Mixed Methods in Unfractured and Fractured Media, *Water Resources Research* (2005), 41, W11412, doi:10.1029/2005WR004339.
8. Huyakorn, P. S., Pinder, G. F.: Computational Methods in Subsurface Flow (1983), Academic Press, Inc., New York.
9. Leveque, R. J.: Finite Volume Methods for Hyperbolic Problems (2002), Cambridge University Press, Cambridge.
10. Lohrenz, J., Bray, B. G., Clark, C. R.: Calculating Viscosities of Reservoir Fluids From Their Compositions, *Journal of Petroleum Technology* Oct. (1964) 1171–1176.
11. Mikyška, J., Firoozabadi, A.: Implementation of higher-order methods for robust and efficient compositional simulation, *Journal of Computational Physics* 229 (2010) 2898–2913.
12. Peng, D. Y., Robinson, D. B.: A New Two-Constant Equation of State, *Industrial and Engineering Chemistry: Fundamentals 15* (1976) 59–64.
13. Polívka O., Mikyška J.: Numerical simulation of multicomponent compressible flow in porous medium, *Journal of Math-for-Industry* Vol. 3 (2011C-7), (2011) 53–60.
14. Russel, T. F., Wheeler, M. F.: Finite Element and Finite Difference Methods for Continuous Flows in Porous Media in: *The Mathematics of Reservoir Simulation, Frontiers in Applied Mathematics* (1983) 35–106, SIAM, Philadelphia.
15. Young, L. C., Stephenson, R. E.: A Generalized Compositional Approach for Reservoir Simulation, *Society of Petroleum Engineers Journal*, Vol.: 23, Issue: 5 (1983) 727–742.
16. Zhang, Q., Wu, Z.-L.: Numerical Simulation for Porous Medium Equation by Local Discontinuous Galerkin Finite Element Method, *Journal of Scientific Computing* (2009) 38: 127–148, DOI 10.1007/s10915-008-9223-7.

Numerical Comparison of Unsteady Channel Compressible Flow with Low Inlet Mach Numbers

P. Pořízková, K. Kozel, and J. Horáček

Abstract This study deals with the numerical solution of a 2D unsteady flow of a compressible viscous fluid in a channel for low inlet airflow velocity. The unsteadiness is caused by a prescribed periodic motion of the channel wall. In study three different governing systems of equations are considered – *Full system, Adiabatic system, Iso-energetic system*. Unsteady flow fields for inlet Mach number $M_\infty = 0.012$ and frequency 100 Hz are presented.

1 Introduction

A current challenging question is a mathematical and physical description of the mechanism for transforming the airflow energy in the glottis into the acoustic energy representing the voice source in humans. The voice source signal travels from the glottis to the mouth, exciting the acoustic supraglottal spaces, and becomes modified by acoustic resonance properties of the vocal tract [1].

Acoustic wave propagation in the vocal tract is usually modeled from incompressible flow models separately using linear acoustic perturbation theory, the wave equation for the potential flow [2] or the Light-hill approach on sound generated aerodynamically [3].

In reality, the airflow coming from the lungs causes self-oscillations of the vocal folds, and the glottis completely closes in normal phonation regimes, generating acoustic pressure fluctuations. In this study, the movement of the boundary channel

P. Pořízková (✉)
Faculty of Mechanical Engineering, Department of Technical Mathematics, Czech Technical University in Prague, Karlovo náměstí 13, Prague 2, 121 35, Czech Republic
e-mail: puncocha@marian.fsik.cvut.cz

K. Kozel · J. Horáček
Institute of Thermomechanics AS CR, Dolejškova 5, Prague 8, Czech Republic
e-mail: Karel.Kozel@fs.cvut.cz; jaromirh@it.cas.cz

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_60,

is known, harmonically opening and nearly closing in the narrowest cross-section of the channel, making the investigation of the airflow field in the glottal region possible. For phonation of vowels, the frequencies of the vocal folds oscillations are in the region from cc 82 Hz for bass up to cc 1,170 Hz for soprano in singing voice, the airflow velocity in the trachea is approximately in the range of 0.3–5.2 ms^{-1} taking into account the tracheal diameter in humans in the range 14.5–17.6 mm [2].

2 Mathematical Models

To describe the unsteady flow of a compressible viscous fluid in a channel, the 2D system of Navier-Stokes equations was considered as a mathematical model. The Navier-Stokes equations were transformed to non-dimensional form. The reference variables are inflow variables (marked with the infinity subscript): the speed of sound $\hat{c}_\infty = 343\,\mathrm{ms}^{-1}$, density $\hat{\rho}_\infty = 1.225\,\mathrm{kg\,m}^{-3}$, temperature $\hat{T}_\infty = 293.15\,\mathrm{K}$, dynamic viscosity $\hat{\eta}_\infty = 18 \cdot 10^{-6}\,\mathrm{Pa \cdot s}$ and a reference length $\hat{L}_r = 0.02\,\mathrm{m}$. The system of Navier-Stokes equations is expressed in non-dimensional conservative form [4] as:

$$\frac{\partial \mathbf{W}}{\partial t} + \frac{\partial \mathbf{F}}{\partial x} + \frac{\partial \mathbf{G}}{\partial y} = \frac{1}{\mathrm{Re}}\left(\frac{\partial \mathbf{R}}{\partial x} + \frac{\partial \mathbf{S}}{\partial y}\right). \tag{1}$$

$\mathbf{W}$ is the vector of conservative variables $\mathbf{W} = [\rho, \rho u, \rho v, e]^T$ where ρ denotes density, u and v are the components of the velocity vector and e is the total energy per unit volume. $\mathbf{F}$ and $\mathbf{G}$ are the vectors of inviscid fluxes and $\mathbf{R}$, $\mathbf{S}$ are the vectors of viscous fluxes. The static pressure p in $\mathbf{F}$ and $\mathbf{G}$ is expressed by the state equation in the form

$$p = (\kappa - 1)\left[e - \frac{1}{2}\rho\left(u^2 + v^2\right)\right], \tag{2}$$

where $\kappa = 1.4$ is the ratio of specific heats.

General Reynolds number in (1) is computed from reference variables $\mathrm{Re} = \hat{\rho}_\infty \hat{c}_\infty \hat{L}_r / \hat{\eta}_\infty$. The non-dimensional dynamic viscosity in the dissipative terms is a function of temperature in the form $\eta = (T/T_\infty)^{3/4}$.

The system of Eqs. (1) and (2) is so-called *Full system*. We present two other governing systems of equations based on the Navier-Stokes equations (1), depend on expression of state equation for static pressure p which is depend on energy flow condition in the system. The second governing system is so-called *Adiabatic system*

$$p = \frac{1}{\kappa}\rho^\kappa, \tag{3}$$

and third governing system is so-called *Iso-energetic system*

$$p = \frac{\rho}{\kappa}\left[1 + \frac{\kappa - 1}{2}\left(\frac{\hat{u}_\infty}{\hat{c}_\infty}\right)^2 - \frac{\kappa - 1}{2}(u^2 + v^2)\right]. \tag{4}$$

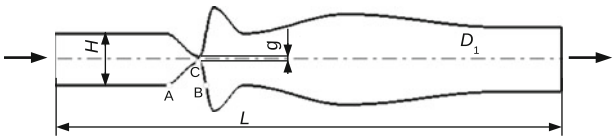


Fig. 1 The computational domain D_1. $L = 8$ (160 mm), $H = 0.8$ (16 mm), $g = 0.08$ (1.6 mm)

Both last systems have the pressure expression independent on variable e and the system of Navier-Stokes equations is reduced on the first three equations.

3 Computational Domain and Boundary Conditions

The bounded computational domain D_1 used for the numerical solution of flow field in the channel is shown in Fig. 1. The domain is symmetric channel, the shape of which is inspired by the shape of the trachea (inlet part), vocal folds, false vocal folds and supraglottal spaces (outlet part) in human vocal tract. The upper and the lower boundaries are the channel walls. A part of the walls changes its shape between the points A and B according to a given function of time and axial coordinate. The gap width is the narrowest part of the channel (in point C) and is oscillating between the minimum $g_{min} = 0.4$ mm and maximum $g_{max} = 2.8$ mm.

The boundary conditions are considered in the following formulation:

1. Upstream conditions: $u_\infty = M_\infty$, $v_\infty = 0$, $\rho_\infty = 1$, p_∞ is extrapolated from D.
2. Downstream conditions: $p_2 = 1/\kappa$ and $(\rho, \rho u, \rho v)$ are extrapolated from D.
3. Flow on the wall: $(u, v) = (u_{wall}, v_{wall})$ and furthermore for *Full system* $\frac{\partial T}{\partial \mathbf{n}} = 0$ $(T = \kappa p/\rho)$.

The general Reynolds number in (1) is multiply with non-dimensional value $M_\infty H$ represents kinematic viscosity scale and for computation of the real problem inlet Reynolds number $\mathrm{Re}_\infty = \hat{\rho}_\infty \hat{c}_\infty M_\infty H \hat{L}_r / \hat{\eta}_\infty$ is used.

4 Numerical Solution

The numerical solution uses finite volume method (FVM) in cell centered form on the grid of quadrilateral cells. In the time-changing domain, the integral form of FVM is derived using the ALE formulation [5].

The explicit predictor-corrector MacCormack (MC) scheme in the domain with a moving grid of quadrilateral cells is used. The scheme is the second order accurate in time and space [4]. To stabilize computation the Jameson artificial dissipation is added to the MC scheme [6].

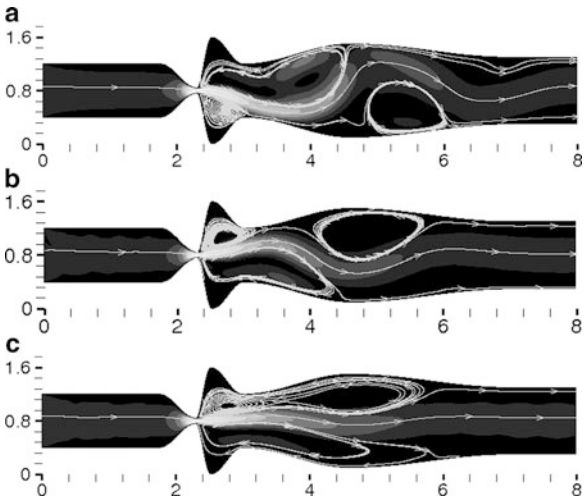


Fig. 2 Initial conditions in domain D_1 computed with all systems. $M_\infty = 0.012$, $\mathrm{Re}_\infty = 4{,}481$, $p_2 = 1/\kappa$, mesh: 450×100. Results are mapped by iso-lines of ratio velocity and by streamlines. (**a**) Full system, $M_{\max} = 0.177$. (**b**) Adiabatic system, $M_{\max} = 0.166$. (**c**) Iso-energetic system, $M_{\max} = 0.160$

The grid used in the channel has successive refinement cells near the wall (see [7]). The minimum cell size in y-direction is $\Delta y_{min} \approx 1/\sqrt{\mathrm{Re}_\infty}$ to capture the boundary layer effects.

5 Numerical Results

The numerical results were obtained (using a specifically developed program) for the following input data: uniform inflow ratio velocity $\frac{\hat{u}_\infty}{\hat{c}_\infty} = M_\infty = 0.012$ ($\hat{u}_\infty = 4.116\,\mathrm{ms}^{-1}$), Reynolds number $\mathrm{Re}_\infty = 4{,}481$, atmospheric pressure $p_2 = 1/\kappa$ ($\hat{p}_2 = 102{,}942\,\mathrm{Pa}$) at the outlet and wall oscillation frequency $\hat{f} = 100\,\mathrm{Hz}$. The computational domain contained 450×100 cells.

The computation has been carried out in two stages. First, a numerical solution is obtained, when the channel between points A and B has a rigid wall fixed in the middle position of the gap width. Then this solution is used as the initial condition for the unsteady simulation.

Figure 2 shows initial conditions of the flows in domain D_1 computed with all systems. The pictures display non-symmetric flow developed behind the narrowest channel cross-section. Figure 3 shows the convergences to the steady state solution for all systems computed using the L_2 norm of momentum residuals (ρu). The graphs indicates the non-stationary solution which is caused probably by eddies separated behind gap and floating away. The worst residuals has numerical solution computed by *Iso-energetic system*.

The numerical simulations of the air-flows computed in domain D_1 with the systems are presented in Figs. 4–6 showing the unsteady flow fields in five time

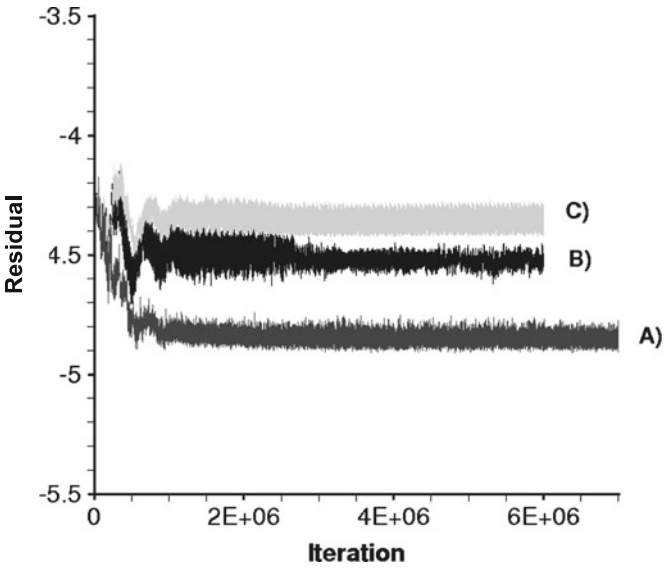


Fig. 3 Convergence to the steady state solution. Computed in domain D_1 with: (*A*) *Full system*, (*B*) *Adiabatic system*, (*C*) *Iso-energetic system*

instants during one vibration period (in the fourth cycle of the wall oscillation). The highest absolute maximum velocity ratio during one vibration period is computed by *Full system* (Fig. 4) where $M_{\max} = 0.535$ ($\hat{u}_{max} = 183.5\,\mathrm{ms}^{-1}$) at g = 1.002 mm (opening phase). *Adiabatic system* and *Iso-energetic system* (Figs. 5 and 6) have same absolute maximum velocity ratio during one vibration period $M_{\max} = 0.199$ ($\hat{u}_{max} = 68.2\,\mathrm{ms}^{-1}$) at $g = 0.993$ mm and $g = 1.09$ mm respectively, during closing phase.

6 Discussion and Conclusions

Authors tested three systems of governing equations for viscous fluid based on Navier-Stokes equation for laminar flow. Numerical solutions showed similar pattern of the flow fields computed with *Full*, Adiabatic and Iso-energetic systems. In numerical solutions with those systems was possible to detect a "Coandă phenomenon" in the flow field patterns. The Coandă effect (the direction of the jet) is independent on the coarseness of mesh but depends on the geometry of the channel, on the type of mesh in the domain, on the computational scheme [8] and on mathematical model of flow. A similar generation of large-scale vortices, vortex convection and diffusion, jet flapping, and general flow patterns were experimentally obtained in physical models of the vocal folds by using PIV (Particle Image Velocimetry) method in [9].

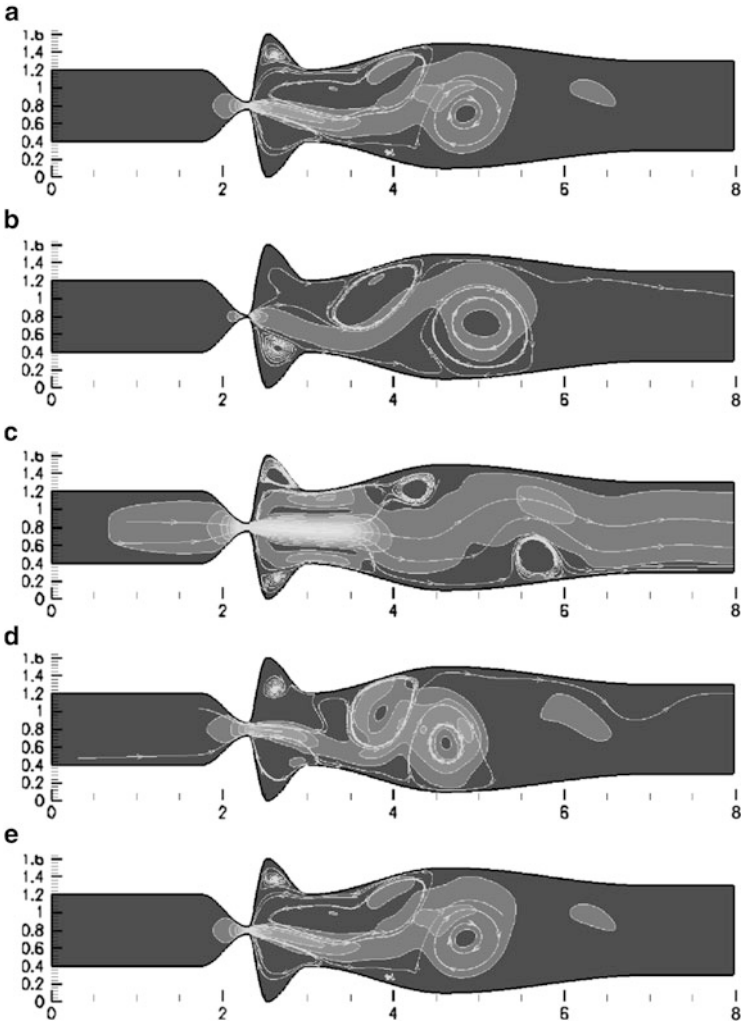


Fig. 4 The unsteady numerical solution of the airflow in D with *Full system* – $\hat{f} = 100\,\text{Hz}$, $M_\infty = 0.012$, $\text{Re}_\infty = 4{,}481$, $p_2 = 1/\kappa$, 450×100 cells. Data computed during the fourth oscillation cycle. Results are mapped by iso-lines of velocity ratio and by streamlines. (**a**) $t = 30\,\text{ms}$, $g = 1.6\,\text{mm}$, $M_{\max} = 0.162$ ($55.6\,\text{ms}^{-1}$). (**b**) $t = 32.5\,\text{ms}$, $g = 0.4\,\text{mm}$, $M_{\max} = 0.236$ ($80.9\,\text{ms}^{-1}$). (**c**) $t = 35\,\text{ms}$, $g = 1.6\,\text{mm}$, $M_{\max} = 0.370$ ($126.9\,\text{ms}^{-1}$). (**d**) $t = 37.5\,\text{ms}$, $g = 2.8\,\text{mm}$, $M_{\max} = 0.097$ ($33.3\,\text{ms}^{-1}$). (**e**) $t = 40\,\text{ms}$, $g = 1.6\,\text{mm}$, $M_{\max} = 0.162$ ($55.6\,\text{ms}^{-1}$)

The computations show that some numerical results of viscous flow in a symmetric channel using a symmetric grid and scheme can be non-symmetrical, depending on the geometry and the Reynolds number (for more see [10]).

The influence of the channel length and frequency of the walls computed with the *Full system* are tested in [7, 11]. The numerical solution showed more streamlined

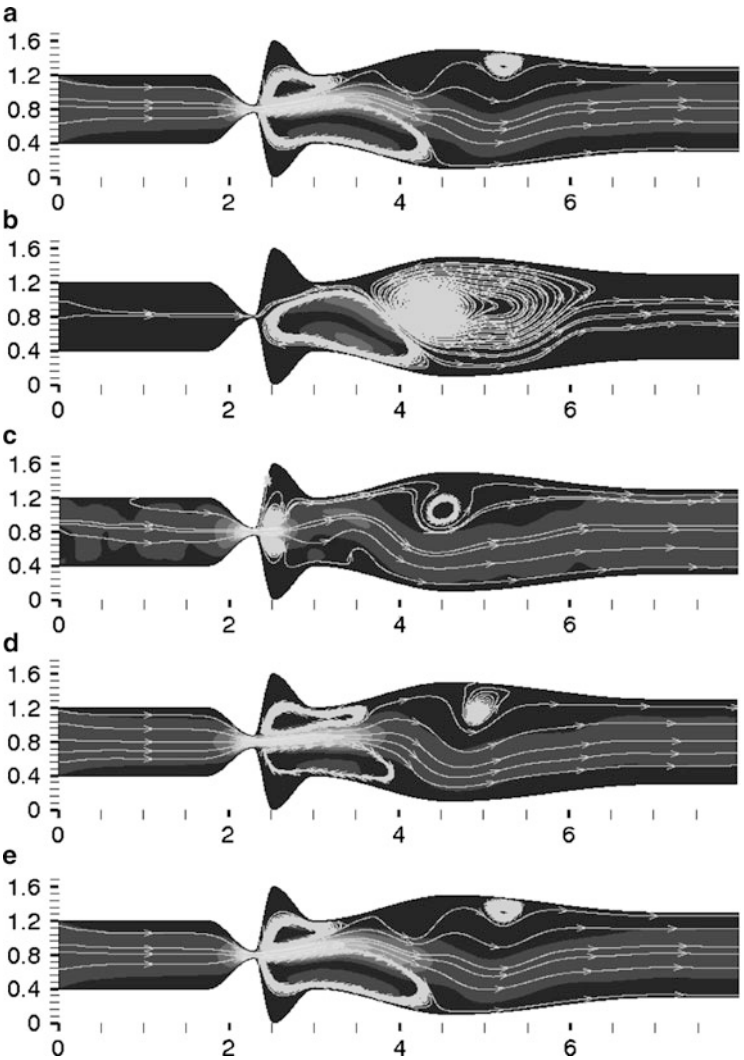


Fig. 5 The unsteady numerical solution of the airflow in D with *Adiabatic system* – $\hat{f} = 100\,\mathrm{Hz}$, $M_\infty = 0.012$, $\mathrm{Re}_\infty = 4{,}481$, $p_2 = 1/\kappa$, 450×100 cells. Data computed during the fourth oscillation cycle. Results are mapped by iso-lines of velocity ratio and by streamlines. (**a**) $t = 30\,\mathrm{ms}$, $g = 1.6\,\mathrm{mm}$, $M_{\max} = 0.167$ ($57.2\,\mathrm{ms}^{-1}$). (**b**) $t = 32.5\,\mathrm{ms}$, $g = 0.4\,\mathrm{mm}$, $M_{\max} = 0.032$ ($10.9\,\mathrm{ms}^{-1}$). (**c**) $t = 35\,\mathrm{ms}$, $g = 1.6\,\mathrm{mm}$, $M_{\max} = 0.171$ ($58.6\,\mathrm{ms}^{-1}$). (**d**) $t = 37.5\,\mathrm{ms}$, $g = 2.8\,\mathrm{mm}$, $M_{\max} = 0.093$ ($31.9\,\mathrm{ms}^{-1}$). (**e**) $t = 40\,\mathrm{ms}$, $g = 1.6\,\mathrm{mm}$, $M_{\max} = 0.167$ ($57.2\,\mathrm{ms}^{-1}$)

flow pattern due to the prolonged outlet part of the channel. The prolongation of the channel inlet part resulted in the developed velocity profile before entering the narrowest channel cross-section.

Next time authors consider use extension to 3D case.

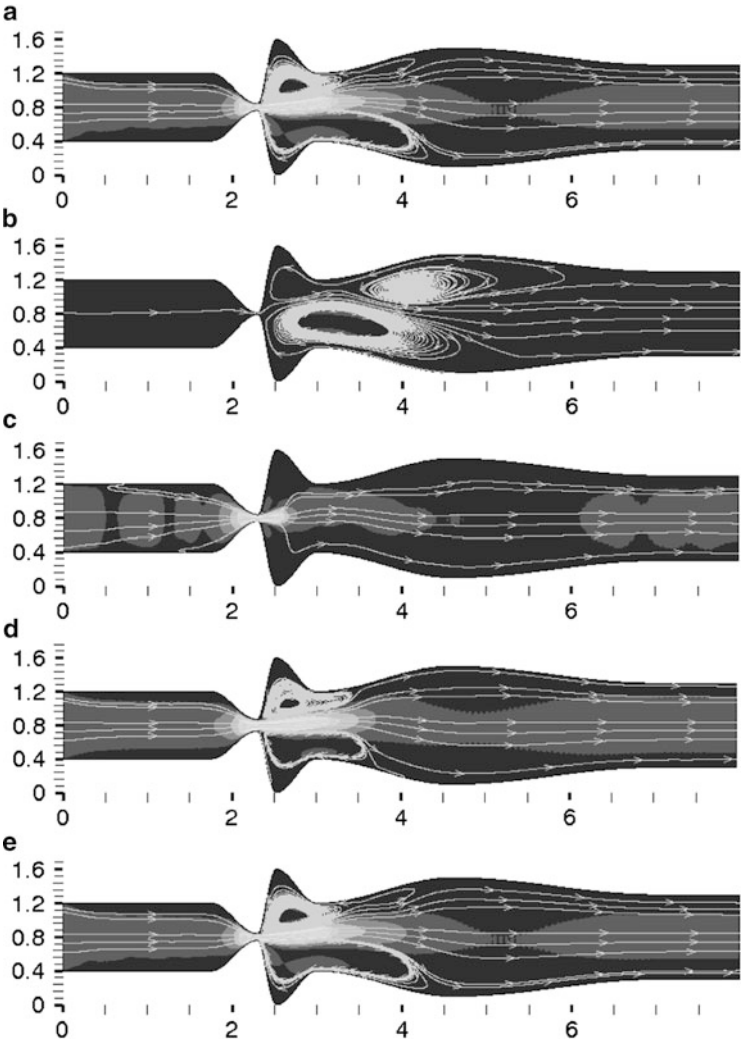


Fig. 6 The unsteady numerical solution of the airflow in D with *Iso-energetic system* – $\hat{f} = 100$ Hz, $M_\infty = 0.012$, $\mathrm{Re}_\infty = 4{,}481$, $p_2 = 1/\kappa$, 450×100 cells. Data computed during the fourth oscillation cycle. Results are mapped by iso-lines of velocity ratio and by streamlines. (**a**) $t = 30\,\mathrm{ms}$, $g = 1.6\,\mathrm{mm}$, $M_{\max} = 0.163$ ($55.9\,\mathrm{ms}^{-1}$). (**b**) $t = 32.5\,\mathrm{ms}$, $g = 0.4\,\mathrm{mm}$, $M_{\max} = 0.030$ ($10.3\,\mathrm{ms}^{-1}$). (**c**) $t = 35\,\mathrm{ms}$, $g = 1.6\,\mathrm{mm}$, $M_{\max} = 0.159$ ($54.5\,\mathrm{ms}^{-1}$). (**d**) $t = 37.5\,\mathrm{ms}$, $g = 2.8\,\mathrm{mm}$, $M_{\max} = 0.097$ ($33.3\,\mathrm{ms}^{-1}$). (**e**) $t = 40\,\mathrm{ms}$, $g = 1.6\,\mathrm{mm}$, $M_{\max} = 0.163$ ($55.9\,\mathrm{ms}^{-1}$)

Acknowledgements This contribution was partially supported by Research Plans MSM 6840770010, GAČR P101/11/0207 and 201/08/0012.

References

1. Titze, I. R., Principles of Voice Production. National Center for Voice and Speech, Iowa City, 2000. ISBN 0-87414-122-2.
2. Titze, I.R., *The Myoelastic Aerodynamic Theory of Phonation*. National Center for Voice and Speech, Iowa City, 2006. ISBN 0-87414-122-2.
3. Zöner, S. and Kalteenbacher, M. and Mattheus, W. and Brücker, C., *Human phonation analysis by 3d aero-acoustic computation*. In: Proceedings of the International Conference on Acoustic NAG/DAGA 2009, 1730–1732, Rotterdam.
4. Fürst, J., Janda, M., Kozel, K., *Finite volume solution of 2D and 3D Euler and Navier-Stokes equations*. In: P. Neustupa, J. & Penel (Eds.), Mathematical fluid mechanics, 173–194, Berlin, 2001.
5. Honzátko, R., Horáček, J., Kozel, K., *Solution of inviscid incompressible flow over a vibrating profile*. In: M. Beneš & M. Kimura & T. Nataki (Eds.), COE Lecture notes, **3**: 26–32, Kyushu university, 2006. ISSN 1881–4042.
6. Jameson, A., Schmidt, W., Turkel, E., *Numerical solution of the Euler equations by the finite volume methods using Runge-Kutta time-stepping schemes*, AIAA, 81–125, 1981.
7. Punčochářová - Pořízková, P., Horáček, J., Kozel, K., Fürst, J., *Numerical simulation of unsteady compressible low Mach number flow in a channel*, Engineering mechanics, **17**(2), 83–97, 2010.
8. Pořízková, P., Kozel, K., Horáček, J., Numerical tests of flow in human vocal tract. In: Interaction of Dynamic Systems with Surroundings and Systems with Feedbacks. Prague: Institute of Thermomechanics, AS CR, v.v.i., 2010, p. 79–86. ISBN 978-80-87012-29-1.
9. Horáček, J., Šidlof, P., Uruba, V., Veselý, J., Radolf, V., Bula, V., *PIV Measurement of Flow-Patterns in Human Vocal Tract Model*. In: Proceedings of the International Conference on Acoustic NAG/DAGA 2009, 1737–1740, Rotterdam, 2009.
10. Punčochářová, P., Kozel, K., Horáček, J., Fürst, J., *An Unsteady Numerical Solution of Viscous Compressible Flows in a Channel for Low Mach Numbers*. Journal of Computational and Applied Mechanics, **8**(2): 175–191, 2007. ISSN 1586-2070.
11. Punčochářová - Pořízková, P., Kozel, K., Horáček, J., Fürst, J., Numerical solutions of unsteady flows with low inlet Mach numbers. Mathematics and computers in simulation, **80**(8): 1795–1805, 2010, Elsevier. ISSN 0378-4754.

Numerical Simulation of Generalized Newtonian and Oldroyd-B Fluids

V. Prokop and K. Kozel

Abstract This paper is dealing with numerical simulation of generalized Newtonian and generalized Oldroyd-B fluids. The Newtonian model of a fluid cannot capture all the phenomena in many fluids with complex microstructure, such as polymers, suspensions and granular materials. The motion of polymeric fluids is described by the conservation of mass and momentum. One shall assume that the fluid is incompressible and temperature variations are negligible. When one considers viscoelastic behavior of polymeric fluids, the extra stress tensor depends not only on the current motion of the fluid, but also on the history of the motion. In this case the extra stress tensor is decomposed into its Newtonian part and its elastic part. Components of the elastic part of the extra stress tensor are computed using the Oldroyd-B constitutive equation. Numerical solution of the arising system of equations is solved using the artificial compressibility method, finite volume method and Runge-Kutta method. Numerical methods are tested in the geometry of constricted channel.

1 Mathematical Model

The motion of polymeric fluids is described by the conservation of mass and momentum [3]. One shall assume that the fluid is incompressible and temperature variations are negligible:

$$\mathrm{div}\mathbf{u} = 0, \tag{1}$$

$$\rho\left(\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u}\cdot\nabla)\mathbf{u}\right) = \mathrm{div}\mathbf{T} - \nabla p, \tag{2}$$

V. Prokop (✉) · K. Kozel
CTU Prague, Karlovo nám. 13, Prague 2, 12000, Czech Republic
e-mail: Vladimir.Prokop@fsi.cvut.cz; Karel.Kozel@fsi.cvut.cz

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_61,

where $\mathbf{u}$ is the velocity vector, ρ is the density, $\mathbf{T}$ is the extra stress tensor, p is an isotropic pressure. The left side of the Eq. (2) represents the force of inertia. This term is important in Newtonian fluid mechanics, but it is often negligible in polymeric fluids. The stress tensor $\mathbf{T}$ represents the force which the material develops in response to being deformed [11].

1.1 Constitutive Models

The mathematical model is complete when a constitutive law, relating $\mathbf{T}$ to the motion, is prescribed. In the case of Newtonian fluids the extra stress tensor $\mathbf{T}$ is proportional to $\mathbf{D}$, the symmetric part of the velocity gradient:

$$\mathbf{T} = 2\mu\mathbf{D} \tag{3}$$

with the viscosity μ being the constant of proportionality and $\mathbf{D}$ is the symmetric part of the velocity gradient:

$$D_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right) \tag{4}$$

If generalized Newtonian fluids are considered, the constant viscosity is replaced with the viscosity function μ_s depending on the shear rate (modified Cross Model)[5, 12]:

$$\mu_s = \mu_s(\dot{\gamma}) = \mu_\infty - \frac{\mu_0 - \mu_\infty}{(1 + (\alpha\dot{\gamma})^b)^a}. \tag{5}$$

The shear rate $\dot{\gamma}$ is defined using the symmetric part of the velocity gradient $\mathbf{D}$:

$$\dot{\gamma} \equiv \sqrt{2\mathrm{tr}(\mathbf{D}^2)}, \tag{6}$$

where $\mu_0 = \lim_{\dot{\gamma}\to 0}\mu(\dot{\gamma})$ and $\mu_\infty = \lim_{\dot{\gamma}\to\infty}\mu(\dot{\gamma})$ are asymptotic viscosity values at zero and infinite shear rates, [2]. The constants of the model are gained from experiment [7]. When one considers viscoelastic behavior of polymeric fluids, the extra stress tensor depend not only on the current motion of the fluid, but also on the history of the motion [11]. In this case the extra stress tensor $\mathbf{T}$ is decomposed into its Newtonian part $\mathbf{T}_s$ and its elastic part $\mathbf{T}_e$. Components of the elastic part of the extra stress tensor are computed using the Oldroyd-B constitutive equation [10]:

$$\frac{\partial\mathbf{T}_e}{\partial t} + (\mathbf{u}\cdot\nabla)\mathbf{T}_e = \frac{2\mu_e}{\lambda_1}\mathbf{D} - \frac{1}{\lambda_1}\mathbf{T}_e + (\mathbf{W}\mathbf{T}_e + \mathbf{T}_e\mathbf{W}) + (\mathbf{D}\mathbf{T}_e + \mathbf{T}_e\mathbf{D}), \tag{7}$$

where tensors $\mathbf{D}$ and $\mathbf{W}$ are the symmetric and antisymmetric parts of the velocity gradient, λ_1 is the relaxation time.

2 Numerical Solution

The resulting systems of equations for pressure, velocity components and components of the elastic part of the extra stress tensor are as follows:

$$u_x + v_y = 0 \tag{8}$$

$$u_t + (u^2 + \bar{p})_x + (uv)_y = \frac{1}{\rho}((\mu_s u_x + t_1)_x + (\mu_s u_y + t_2)_y) \tag{9}$$

$$v_t + (uv)_x + (v^2 + \bar{p})_y + = \frac{1}{\rho}((\mu_s v_x + t_2)_x + (\mu_s v_y + t_3)_y) \tag{10}$$

where $\bar{p} = \frac{p}{\rho}$ is the kinematic pressure and $t_i, i = 1, 2, 3$ are the components of the symmetric tensor $\mathbf{T}_e$ (the elastic part of the extra stress tensor). The Oldroyd-B constitutive equations is written in the component form:

$$(t_1)_t + (ut_1)_x + (vt_t)_y = 2\frac{\mu_e}{\lambda_1}u_x - \frac{1}{\lambda_1}t_1 + 2(t_1 u_x + t_2 u_y) \tag{11}$$

$$(t_2)_t + (ut_2)_x + (vt_2)_y = \frac{\mu_e}{\lambda_1}(u_y + v_x) - \frac{1}{\lambda_1}t_2 + t_3 u_y + t_1 v_x + t_2(u_x + v_y) \tag{12}$$

$$(t_3)_t + (ut_3)_x + (vt_3)_y = 2\frac{\mu_e}{\lambda_1}v_y - \frac{1}{\lambda_1}t_3 + 2(t_2 v_x + t_3 v_y) \tag{13}$$

The preceeding system (8)–(10) of equations is solved using the artificial compressibility method [4]. For discretization of spatial derivatives of both systems of Eqs. (8)–(10) and (11)–(13) a finite volume method is employed [6, 8]. The inviscid fluxes are discretized in the central manner, the viscous fluxes are computed using dual finite volume cells of diamond type, [1, 9]. The resulting system of ODEs is solved using three-stage Runge-Kutta method [6] with steady boundary conditions. The second system of Eqs. (11)–(13) is stabilized with the term proportional to the second order derivative of $\mathbf{T}_e$. The Dirichlet boundary conditions for the velocity $\mathbf{u}$ and the elastic part of the extra stress tensor $\mathbf{T}_e$ are prescribed at the inlet. At the outlet, the Dirichlet boundary condition is imposed on the pressure p and the Neumann condition holds for $\mathbf{T}_e$. On the wall, there are prescribed the no-slip condition for $\mathbf{u}$ and the Neumann condition for $\mathbf{T}_e$, [2].

3 Numerical Results

In this section are presented numerical results for four different flow settings in the geometry of a constricted channel. The following values of model parameters, taken from the literature [7], are used in the following computations:

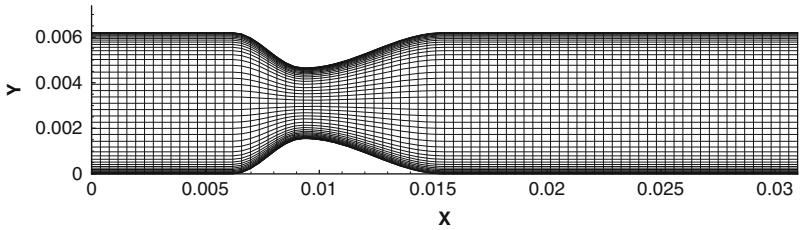


Fig. 1 Mesh of the constricted channel: 80 × 42 cells

Table 1 Computed cases

Newtonian	Generalized Newtonian	Oldroyd-B	Generalized Oldroyd-B
$\mu_s(\dot{\gamma}) = \mu_\infty$	$\mu_s(\dot{\gamma})$	$\mu_s(\dot{\gamma}) = \mu_\infty$	$\mu_s(\dot{\gamma})$
$\mathbf{T}_e = 0$	$\mathbf{T}_e = 0$	$\mathbf{T}_e$	$\mathbf{T}_e$

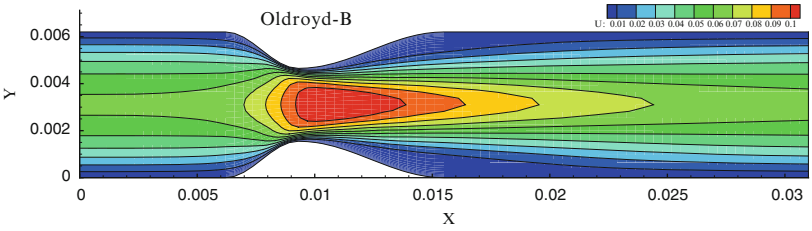


Fig. 2 Oldroyd-B fluid: velocity field

- Parameters of the modified Cross Model, [7]:
  - $\mu_0 = 0.16$ Pa.s, $\mu_\infty = 0.0036$ Pa.s,
  - $a = 1.23$, $b = 0.64$, $\alpha = 8.2$ s.
- Relaxation time: $\lambda_1 = 0.06$ s
- Inlet velocity: $U_0 = 0.0615\,\mathrm{m}\cdot\mathrm{s}^{-1}$
- Width of the channel: $L_0 = 2R = 0.0062$ m
- Length of the channel: $L_c = 10R = 0.031$ m
- Stenosis (Constriction): 50
- Density: $\rho = 1,050\,\mathrm{kg}\cdot\mathrm{m}^{-3}$

The first Fig. 1 shows the geometry and the mesh of the constricted channel.

In the Table 1 are written the parameters of four settings of the considered computations:

The results on the Figs. 2–4, and 9 shows that for shear-thinning flows (Generalized Oldroyd-B fluid, Generalized Newtonian fluid) the recirculation zone becomes shorter. The characteristic viscosity in Newtonian case is μ_∞ and thus shear-thinning viscosity of the form of modified Cross model leads to the increase of the local viscosity in the low shear regions (Fig. 5).

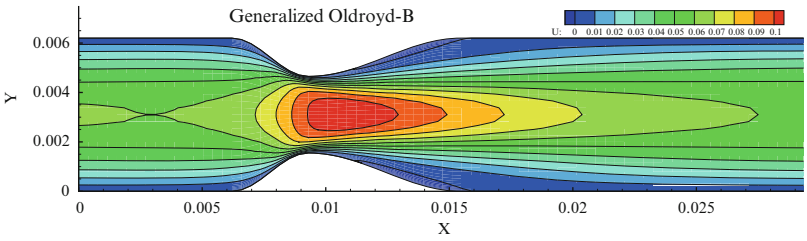


Fig. 3 Generalized Oldroyd-B fluid: velocity field

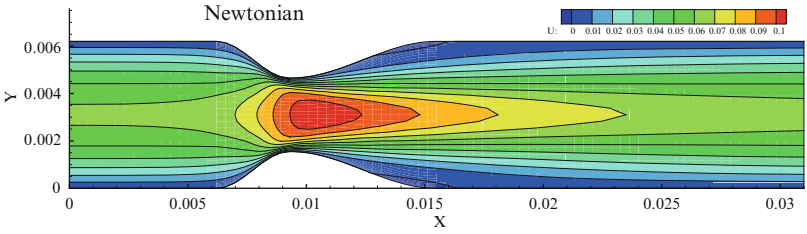


Fig. 4 Newtonian fluid: velocity field

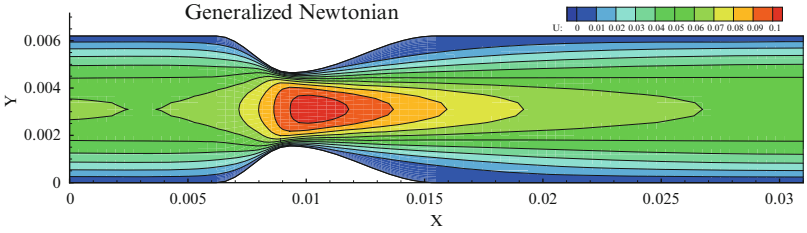


Fig. 5 Generalized Newtonian fluid: velocity field

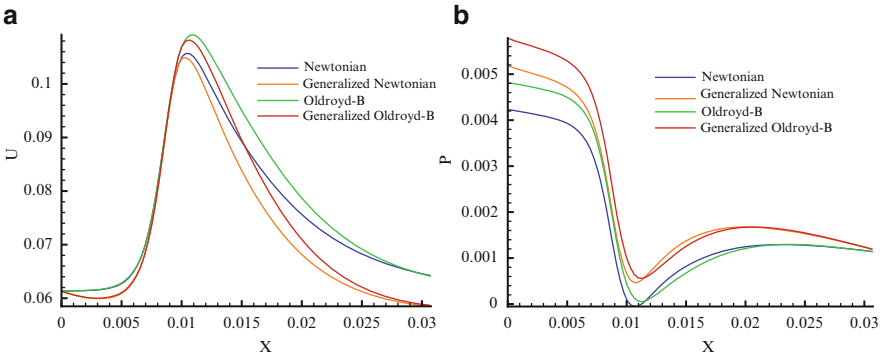


Fig. 6 Comparison of (**a**) velocity and (**b**) pressure distributions along the axis

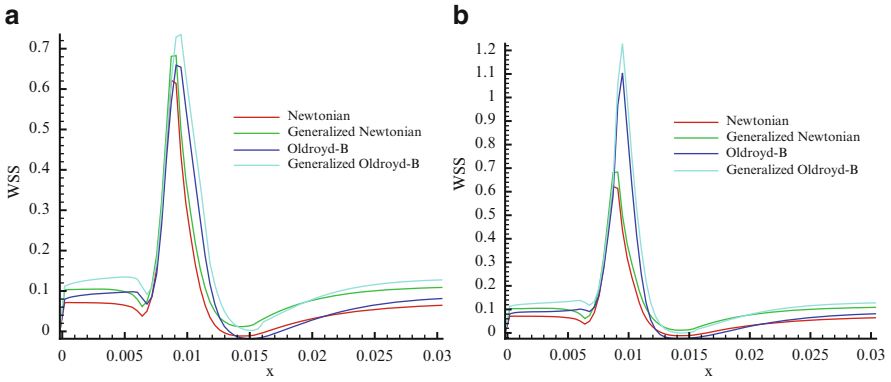


Fig. 7 WSS comparison (**a**) $e_t = 0.001$ (**b**) $e_t = 0.0001$

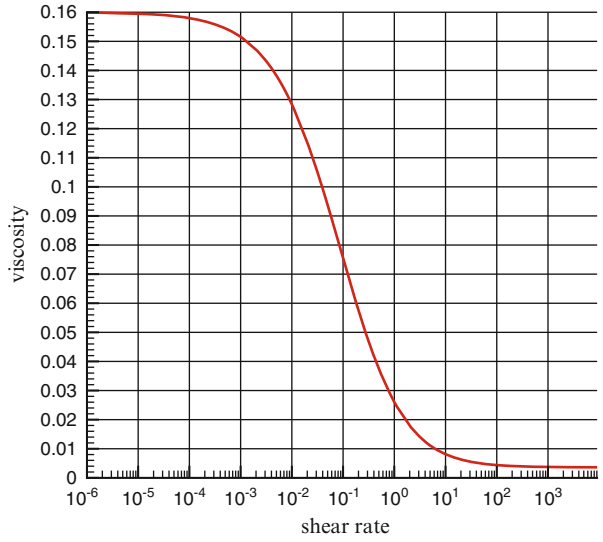


Fig. 8 Viscosity function

The axial pressure profiles for all four models show that the main effect of shear-thinning behavior is visible mainly in the recirculation zone, where the local viscosity and thus the resistance to the flow increases significantly, see Fig. 6. The effects of viscoelasticity are about one order of magnitude lower in this case.

Another important parameter in the flow is the wall shear stress (WSS). It represents the tangential component of the surface force at the wall, acting against the fluid flow. The negative values of the WSS can be found in regions of reversal flow and also in regions where the flow speeds up significantly. In Fig. 7, there is a comparison of the wall shear stress in four different cases and also for two different

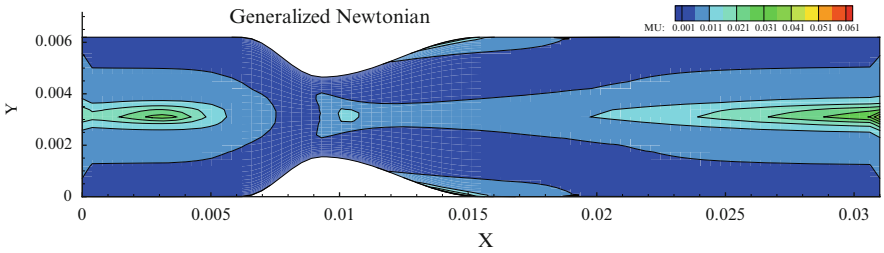


Fig. 9 Viscosity distribution: generalized Newtonian fluid

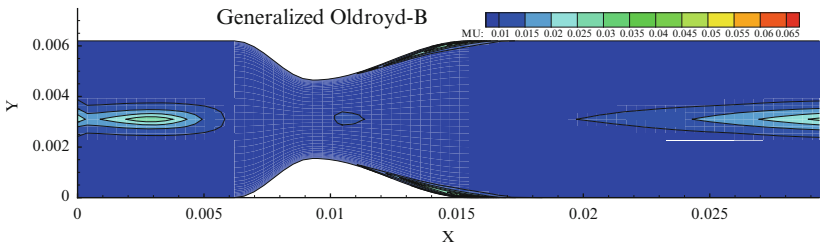


Fig. 10 Viscosity distribution: generalized Oldroyd-B fluid

constants e_t of numerical stabilization of the extra stress tensor $\mathbf{T_e}$. It is visible that e_t significantly affects the peak values of the WSS.

The viscosity distribution according to the modified Cross Model and previously described parameters is shown in the Fig. 8.

The two last Figs. 9 and 10 show variation in the viscosity distribution for generalized Newtonian and generalized Oldroyd-B fluids.

3.1 Conclusion

Four different types of fluid models were implemented and tested in the geometry of the constricted channel. The results show that in the presented cases of flows the effect of the shear-thinning viscosity was much stronger than the phenomena of viscoelasticity. These effects are strongly dependent on the geometry and there are plans to perform computations in the geometry of the stenotic channel and bypass. There were done similar computations with slightly different numerical methods by other researchers, but this work is basis for further development and extension into three dimensions and complicated geometries.

Acknowledgements This work was sponsored by Research Plan MSM 6840770010, GA AS CR No. IAA 100190804, GA CR No. 101/09/1539.

References

1. J. Blazek: Computational Fluid Dynamics: Principles and Applications, Elsevier, 2001
2. T. Bodnar, A. Sequeira, M. Prosi: On the shear-thinning and viscoelastic effects of blood flow under various flow rates, Applied Mathematics and Computation 217, pp. 5055–5067, 2011
3. M. Brdička a kolektiv: Mechanics of Continuum (in Czech), Academia, 2000
4. A. J. Chorin: A Numerical method for solving incompressible viscous flow problems, Journal of Computational Physics 135 (1997), 118–125
5. G. P. Galdi, R. Rannacher, A. M. Robertson, S. Turek: Hemodynamical flows, modeling, analysis and simulation, Birkhauser Verlag AG, Basel Switzerland, 2008
6. Ch. Hirsch: Numerical Computation of Internal and External Flows, Butterworth-Heinemann, 2nd edition, 2007
7. A. Leuprect, K. Perktold: Computer simulation of non-Newtonian effects of blood flow in large arteries, Computer Methods in Biomechanics and Biomechanical Engineering 4, 2001, pp. 149–163
8. R. J. LeVeque: Numerical methods for conservation laws, Birkhauser Verlag, Basel, Switzerland, 1990
9. J. Fořt, K. Kozel, P. Louda, J. Fürst: Numerical methods solving the flow problems III (in Czech), CTU Prague, 2004
10. R. G. Owens, T. N. Phillips: Computational Rheology, Imperial College Press, 2002
11. M. Renardy: Mathematical analysis of viscoelastic flows, SIAM, 2000
12. A. Sequeira, J. Janela: An overview of some mathematical models of blood rheology, M. S. Pereira (ed.), A portrait of state-of-the-art research at the Technical University of Lisbon, 65–87, Springer, 2007

Layer-Adapted Meshes Versus Weak Dirichlet Conditions in Low-Turbulent Flow Simulation

L. Röhe and G. Lube

Abstract For a family of variational multiscale methods we perform an a-priori error analysis for inf-sup stable finite element pairs in low-turbulent incompressible flow problems. This is done for underlying layer-adapted meshes with strong Dirichlet boundary conditions and for isotropic meshes with weak Dirichlet boundary conditions. For both approaches we provide first numerical results in a three-dimensional channel at $Re_\tau = 180$.

1 Introduction

We apply a projection-based variational multiscale (VMS) method, originally published in [5], to the simulation of low-turbulent, wall-bounded, incompressible flow. Our approach relies on inf-sup stable finite element pairs for velocity/pressure. The (semidiscrete) a-priori analysis of this method, given in [7], allows rather general nonlinear, piecewise constant coefficients of the subgrid models for the unresolved scales. The analysis takes advantage of divergence preserving interpolation which had been considered in [3] for the case of simplicial isotropic meshes. An extension to anisotropic quadrilateral meshes has been recently considered in [2].

Here we consider two approaches: (i) layer-adapted anisotropic quadrilateral grids (see, e.g., [4]), and (ii) weakly-enforced Dirichlet boundary conditions on isotropic meshes (see, e.g., [1]). For the coefficients of the projection-based VMS method we apply an approach motivated by arguments of the popular Smagorinsky model. Based on the results in [2,7], we show the applicability of the recent a-priori analysis for VMS methods to both approaches. Numerical results for both variants

L. Röhe (✉) · G. Lube
Georg-August University Göttingen, Institute for Numerical and Applied Mathematics, Göttingen, Germany
e-mail: roehe@math.uni-goettingen.de; lube@math.uni-goettingen.de

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_62,

are presented for the well-known benchmark of low-turbulent flow in a three-dimensional channel at $Re_\tau = 180$.

2 Variational Multiscale Model of Navier-Stokes Problem

Let $\Omega \subset \mathbf{R}^3$ be a bounded polyhedral domain. The incompressible Navier-Stokes equations consists of finding velocity $\mathbf{u}$ and pressure p such that

$$\begin{aligned} \partial_t \mathbf{u} - \nabla \cdot (2\nu \mathbf{Du}) + (\mathbf{u} \cdot \nabla)\mathbf{u} + \nabla p &= \mathbf{f} \qquad \text{in } (0,T] \times \Omega \\ \nabla \cdot \mathbf{u} &= 0 \qquad \text{in } (0,T] \times \Omega \\ \mathbf{u}|_{t=0} &= \mathbf{u}^0 \qquad \text{in } \Omega \end{aligned}$$

with deformation tensor $\mathbf{Du} = \frac{1}{2}(\nabla \mathbf{u} + (\nabla \mathbf{u})^T)$ and viscosity ν. For simplicity, we consider homogeneous Dirichlet boundary conditions and the solution spaces

$$V = [H_0^1(\Omega)]^3, \qquad Q = L^2_*(\Omega) := \{q \in L^2(\Omega) : \int_\Omega q \, dx = 0\}.$$

The variational formulation reads: find $(\mathbf{u}, p)\colon (0,T] \to V \times Q$ s.t. $\forall (\mathbf{v}, q) \in V \times Q$

$$(\partial_t \mathbf{u}, \mathbf{v}) + (2\nu \mathbf{Du}, \mathbf{Dv}) + b_S(\mathbf{u}, \mathbf{u}, \mathbf{v}) - (p, \nabla \cdot \mathbf{v}) + (q, \nabla \cdot \mathbf{u}) = (\mathbf{f}, \mathbf{v})$$

with

$$b_S(\mathbf{u}, \mathbf{v}, \mathbf{w}) := 1/2[((\mathbf{u} \cdot \nabla)\mathbf{v}, \mathbf{w}) - ((\mathbf{u} \cdot \nabla)\mathbf{w}, \mathbf{v})].$$

Let $\mathscr{T}_h$ be an admissible triangulation of Ω such that $\overline{\Omega} = \cup_{K \in \mathscr{T}_h} \overline{K}$. Here, we consider inf-sup stable velocity-pressure finite element (FE) spaces $V_h \times Q_h \subset V \times Q$ with the discrete inf-sup condition

$$\exists \beta \neq \beta(h) \;\; s.t. \;\; \inf_{q_h \in Q_h} \sup_{\mathbf{v}_h \in V_h} \frac{(q_h, \nabla \cdot \mathbf{v_h})}{\|q_h\|_0 \|\nabla \mathbf{v}_h\|_0} \geq \beta > 0. \tag{1}$$

The Galerkin method reads: find $(\mathbf{u}_h, p_h)\colon (0,T] \to V_h \times Q_h$ s.t. $\forall (\mathbf{v}_h, q_h) \in V_h \times Q_h$

$$(\partial_t \mathbf{u}_h, \mathbf{v}_h) + (2\nu \mathbf{Du}_h, \mathbf{Dv}_h) + b_S(\mathbf{u}_h, \mathbf{u}_h, \mathbf{v}_h) - (p_h, \nabla \cdot \mathbf{v}_h) + (q_h, \nabla \cdot \mathbf{u}_h) = (\mathbf{f}, \mathbf{v}_h).$$

The idea of VMS was developed in 1995 by Hughes et al. Its application to scale separation in turbulence modeling started around 2000. The basic aim is to model the influence of smallest (unresolved) scales onto the small scales. Following an idea of Layton [5], we define a coarser FE space L_H for the deformation tensor where

$$\{0\} \subseteq L_H \subseteq L := \{\mathbf{L} = (l_{ij}) : l_{ij} \in L^2(\Omega) \forall i, j \in \{1,2,3\}\}.$$

Define the L^2-orthogonal projection operator $P_H : L \to L_H$ and the small scales via $\kappa(\mathbf{Du}_h) := \mathbf{Du}_h - P_H(\mathbf{Du}_h)$ with fluctuation operator $\kappa := Id - P_H$. Then the VMS method reads: find $(\mathbf{u}_h, p_h)\colon (0,T] \to V_h \times Q_h$ s.t. $\forall (\mathbf{v}_h, q_h) \in V_h \times Q_h$:

$$(\partial_t \mathbf{u}_h, \mathbf{v}_h) + 2\nu\,(\mathbf{Du}_h, \mathbf{Dv}_h) + \big(\nu_T(\mathbf{u}_h)\kappa\mathbf{Du}_h, \kappa\mathbf{Dv}_h\big) \\ + b_S(\mathbf{u}_h, \mathbf{u}_h, \mathbf{v}_h) - (p_h, \nabla\cdot\mathbf{v}_h) + (\nabla\cdot\mathbf{u}_h, q_h) = (\mathbf{f}, \mathbf{v}_h)$$

with cellwise constant $\nu_T^K(\mathbf{u}_h) := \nu_T(\mathbf{u}_h)|_K \geq 0$ for all $K \in \mathscr{T}_h$. The space

$$V_h^{\mathrm{div}} := \{\mathbf{v}_h \in V_h : (\nabla\cdot\mathbf{v}_h, q_h) = 0\ \forall q_h \in Q_h\}$$

of discretely divergence free functions is not empty thanks to condition (1). Then the discrete problem reduces to: find $\mathbf{u}_h\colon (0,T] \to V_h^{\mathrm{div}}$ s.t. $\forall \mathbf{v}_h \in V_h^{\mathrm{div}}$:

$$(\partial_t \mathbf{u}_h, \mathbf{v}_h) + 2\nu\,(\mathbf{Du}_h, \mathbf{Dv}_h) + \big(\nu_T(\mathbf{u}_h)\kappa\mathbf{Du}_h, \kappa\mathbf{Dv}_h\big) + b_S(\mathbf{u}_h, \mathbf{u}_h, \mathbf{v}_h) = (\mathbf{f}, \mathbf{v}_h) \quad (2)$$

with initial condition $\mathbf{u}_h(0) = I_h\mathbf{u}_0$ and an interpolation operator $I_h : V \to V_h^{\mathrm{div}}$. Let us start with the following stability result given in [7], Lemma 3.1.

Lemma 1. *Let* $\mathbf{f} \in [L^1(0,T;L^2(\Omega))]^3$, $\mathbf{u}_0 \in [L^2(\Omega)]^3$*; then we obtain for* $t \in (0,T]$ *control of kinetic energy and control of dissipation and subgrid terms, respectively:*

$$\|\mathbf{u}_h\|_{L^\infty(0,t;L^2(\Omega))} \leq K(\mathbf{f}, \mathbf{u}_0) \equiv \|\mathbf{u}_0\|_0 + \|\mathbf{f}\|_{L^1(0,t;L^2(\Omega))}$$

$$\nu\|\mathbf{Du}_h\|^2_{L^2(0,t;L^2(\Omega))} + \frac{1}{2}\int_0^t \sum_K \nu_T^K(\mathbf{u}_h)\|\kappa_u\mathbf{Du}_h\|^2_{0,K}\,dt \leq 3K^2(\mathbf{f}, \mathbf{u}_0).$$

One technical trick in the error analysis is to rewrite the turbulence term as

$$\sum_K \nu_T^K(\mathbf{u}_h)\|\kappa_u\mathbf{Dv}_h\|^2_{0,K} = \sum_K \nu_T^K(\mathbf{u}_h)\Big(1 - \frac{\|P_H\mathbf{Dv}_h\|^2_{0,K}}{\|\mathbf{Dv}_h\|^2_{0,K}}\Big)\|\mathbf{Dv}_h\|^2_{0,K}$$

and to set

$$\nu_{\mathrm{mod}}(\mathbf{u}_h, \mathbf{v}_h) := 2\nu + \nu_T^K(\mathbf{u}_h)\Big(1 - \frac{\|P_H\mathbf{Dv}_h\|^2_{0,K}}{\|\mathbf{Dv}_h\|^2_{0,K}}\Big) \geq 2\nu.$$

The following semidiscrete a-priori estimate w.r.t. the mesh-dependent expression

$$\Big|\Big|\mathbf{v}_h\Big|\Big|_t^2 := \frac{1}{2}\|\mathbf{v}_h(t)\|_0^2 + \nu\,\|\mathbf{Dv_h}\|^2_{L^2(0,t;L^2(\Omega))} + \int_0^t \sum_{K\in\mathscr{T}_h} \nu_T^K(\mathbf{u}_h)\,\|\kappa\mathbf{Dv}_h\|^2_{0,K}\,dt$$

is a variant of Theorem 3.5 in [7].

Theorem 1. *Under the assumptions of Lemma 1 and for a sufficiently smooth solution* $\mathbf{u}$ *of the Navier-Stokes model (see Ref. [7]) it holds for the solution* $\mathbf{u}_h$ *of (2)*

$$\Big|\Big|\Big|\mathbf{u}_h - I_h\mathbf{u}\Big|\Big|\Big|_t^2 \leq Ce^{h(t)} \int_0^t g(s)\, ds, \qquad t \in (0,T) \tag{3}$$

with $h(t) := C\,\|\mathbf{Du}(t)\|_0^4/[\min_K \nu_{mod}^K(\mathbf{u}_h(t), \mathbf{e}_h^u(t))]^3$ *and*

$$\begin{aligned} g(t) := \max_K \nu_T^K(\mathbf{u}_h(t))\Big(\|\kappa\mathbf{Du}(t)\|_0^2 + \|\mathbf{D}(\mathbf{u} - I_h\mathbf{u}(t))\|_0^2\Big) \qquad (4)\\ + \frac{1}{\min_K \nu_{mod}^K(\mathbf{u}_h(t), \mathbf{e}_h^u(t))}\Big[\|\partial_t(\mathbf{u} - I_h\mathbf{u})(t)\|_{-1,\Omega}^2 + \inf_{\tilde{p}_h \in Q_h} \|(p - \tilde{p})(t)\|_0^2\\ + \Big(\|\mathbf{Du}(t)\|_0^2 + \|\mathbf{u}_h(t)\|_0\|\mathbf{Du}_h(t)\|_0\Big)\|\mathbf{D}(\mathbf{u} - I_h\mathbf{u})(t)\|_0^2\Big]. \end{aligned}$$

Let us consider Taylor-Hood elements with $V_h \times Q_h = [\mathbf{Q}_k]^3 \times \mathbf{Q}_{k-1}$ with $k \geq 2$ on isotropic meshes $\mathscr{T}_h$. Then we can apply the V_h^{div}-interpolation operator I_h of Girault and Scott [3] and the interpolation properties of the fluctuation operator κ. Under the requirement $\max_K \nu_T^K(\mathbf{u}_h(t)) \leq Ch_K^2$, we obtain from (3)–(4) the estimate

$$\Big|\Big|\Big|\mathbf{u}_h - I_h\mathbf{u}_h\Big|\Big|\Big|_t^2 \leq C(\nu, \nu_T, T, \mathbf{u})h^{2k}, \quad t \in (0,T]. \tag{5}$$

A possible choice is $\nu_T|_K = C_*\Delta^2\|\kappa\mathbf{Du}_h\|_{0,K}/\sqrt{vol(K)}$ for all $K \in \mathscr{T}_h$, user chosen constant C_*, and filter width $\Delta \sim h_K$, see Ref. [7].

3 Applications with Layer-Adapted Meshes

Theorem 1 can be applied to layer-adapted meshes as well. Here we consider tensor-product meshes $\mathscr{T}_h$ in $d \in \{2,3\}$ dimensions where the transformation from reference cell $\hat{K} = (-1,1)^d$ to another cell $K \in \mathscr{T}_h$ can be described by the transformation $x = \mathrm{diag}(h_{1,K}, \ldots, h_{d,K})\,\widehat{x} + a_K$ with the local mesh sizes $h_{i,K}$ into direction i and a shift $a_K \in \mathbf{R}^d$. Assume that the mesh size in direction d is locally the smallest one s.t. $0 < h_{d,K} \leq h_{i,K}\ \forall i \in \{1, \ldots, d-1\}, \forall K \in \mathscr{T}_h$ whereas the mesh sizes in the other directions are isotropic. Moreover, we suppose no abrupt change in the element sizes of neighboring cells, i.e. $h_{i,K'} \leq Ch_{i,K} \leq Ch_{i,K'} \forall K, K' \in \mathscr{T}_h,\ \overline{K'} \cap \overline{K} \neq \emptyset$. Such a construction is considered in Fig. 1 (right).

In [2] we obtained for a divergence-preserving interpolator $I_h : V \to V_h^{\mathrm{div}}$ the following (presumably suboptimal) interpolation estimate:

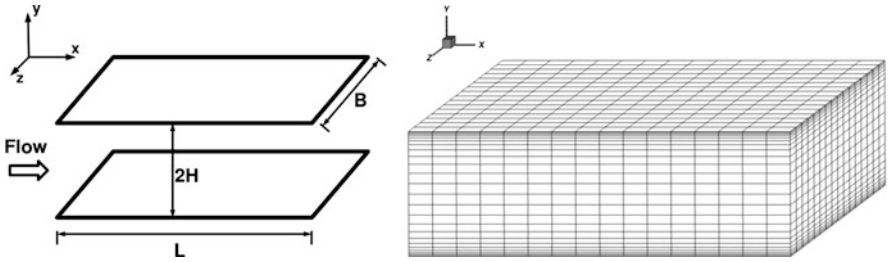


Fig. 1 Channel flow: domain (*left*) and mesh (*right*)

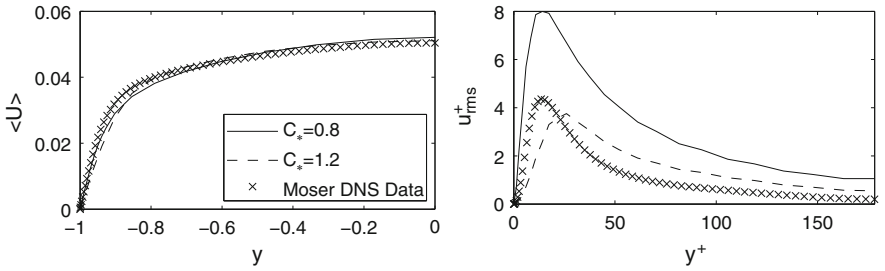


Fig. 2 Channel flow with VMS and $\mathbf{Q}_2/\mathbf{Q}_1$ vs $\mathbf{Q}_0^{disc}$

$$\|\mathbf{v} - I_h\mathbf{v}\|^2_{H^m(K)} \leq C\gamma_i^2 (h_{1,K}/h_{d,K})^2 \sum_{|\alpha|=l-m} h_K^{2\alpha} |D^\alpha \mathbf{v}|^2_{H^m(\omega(K))} \quad \forall \mathbf{v} \in H^l(\omega(K))^d$$

for $m \in \{0, 1\}$. Here γ_i is the maximal aspect ratio of a patch $\omega(K)$ containing K, and $h_{1,K}/h_{d,K}$ is the local aspect ratio of K. The latter estimate can be applied to bound (3)–(4) on layer-adapted meshes.

We now consider a channel flow in $\Omega = (0, H) \times (0, L) \times (0, B)$ with $H = 1$, $L = 4\pi$, $B = \frac{4\pi}{3}$, see Fig. 1 (left). The viscosity $\nu = 1.5 \times 10^{-5}$ corresponds to $Re_\tau = 180$. The initial value for the velocity $\mathbf{u}$ is taken as the known mean profile with random noise. We apply a BDF(2)-scheme with time step $\partial t = 0.86$ and the Taylor-Hood element $\mathbf{Q}_2/\mathbf{Q}_1$ on an anisotropic Cartesian mesh with $16 \times 24 \times 16$ cells and $y(j) = H(\tanh(2(2j/N - 1))/(\tanh(2) + 1),\ j \in \{0, \ldots, N = 24\}$, see Fig. 1 (right).

As reference data from direct numerical simulation we refer to [6]. We are mainly interested in reference values of means, the main channel profile $\langle U \rangle = \lim_{\delta\to\infty} \frac{1}{\delta} \int_{t_0}^{t_0+\delta} U\ dt$ with $U(t, \mathbf{x}) = \mathbf{u}(t, \mathbf{x}) \cdot \mathbf{e}_1$ and the scaled Reynolds stress $u' = U - \langle U \rangle$, $y^+ = yu_\tau/\nu$, see Fig. 2. The numerical results are reasonable on the given coarse mesh $\mathscr{T}_h$. We did not observe a significant influence of the maximal aspect ratio which here is $\frac{1}{16}L/y(1) \approx 55$.

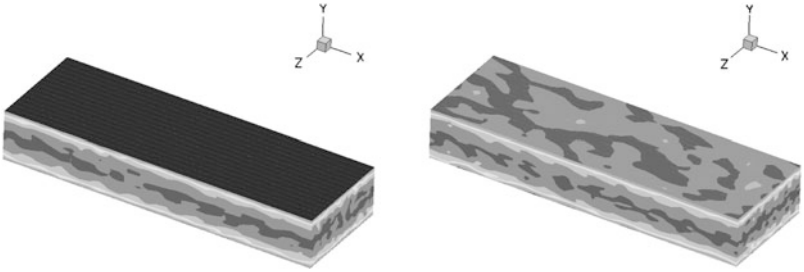


Fig. 3 Strongly (*left*) and weakly (*right*) imposed Dirichlet boundary conditions

4 Applications with Weak Dirichlet Boundary Conditions

An anisotropic mesh refinement becomes more and more expensive with increasing Reynolds number Re_τ. An alternative is to consider weakly enforced Dirichlet conditions on isotropic meshes. Here we follow the framework of Bazilevs et al. [1]. Figure 3 shows the difference of strongly and weakly enforced Dirichlet data.

We assume for simplicity the case of homogeneous Dirichlet data $\mathbf{u} = \mathbf{0}$. Moreover, we divide the boundary $\partial\Omega = \Gamma_{\text{in}} \dot\cup \Gamma_{\text{out}} \dot\cup \Gamma_0$ with $\Gamma_{\text{in}} : \mathbf{u}_h \cdot \mathbf{n} < 0$, $\Gamma_{\text{out}} : \mathbf{u}_h \cdot \mathbf{n} > 0$, and $\Gamma_0 : \mathbf{u}_h \cdot \mathbf{n} = 0$. Following Ref. [1] we consider the modified problem: find $\mathbf{u}_h : (0,T] \to V_h \subset [H^1(\Omega)]^d$, $p_h : (0,T] \to Q_h$ s.t. $\forall (\mathbf{v}_h, q_h) \in V_h \times Q_h$:

$$
\begin{aligned}
&(\partial_t \mathbf{u}_h, \mathbf{v}_h) + 2\nu\,(\mathbf{D}\mathbf{u}_h, \mathbf{D}\mathbf{v}_h) + (\nu_T \kappa \mathbf{D}\mathbf{u}_h, \kappa \mathbf{D}\mathbf{v}_h) + b_S\,(\mathbf{u}_h, \mathbf{u}_h, \mathbf{v}_h) \\
&\qquad + B_{\text{wall}}(\mathbf{u}_h, p_h; \mathbf{v}_h, q_h) - (p_h, \nabla \cdot \mathbf{v}_h) + (\nabla \cdot \mathbf{u}_h, q_h) = (\mathbf{f}, \mathbf{v}_h)
\end{aligned}
$$

with

$$
\begin{aligned}
B_{\text{wall}}(\mathbf{u}_h, p_h; \mathbf{v}_h, q_h) = &-(2\nu \mathbf{D}\mathbf{u}_h \cdot \mathbf{n}, \mathbf{v}_h)_{\partial\Omega} + \left(p_h, \mathbf{v}_h \cdot \mathbf{n}\right)_{\partial\Omega} - \left(q_h, \mathbf{u}_h \cdot \mathbf{n}\right)_{\partial\Omega} \\
&-\frac{1}{2}((\mathbf{u}_h \cdot \mathbf{n})\mathbf{u}_h, \mathbf{v}_h)_{\partial\Omega} - (\mathbf{u}_h, 2\nu \mathbf{D}\mathbf{v}_h \cdot \mathbf{n})_{\partial\Omega} + \frac{1}{2}(\mathbf{u}_h, (\mathbf{u}_h \cdot \mathbf{n})\mathbf{v}_h)_{\Gamma_{\text{out}}} \\
&+ (\mathbf{u}_h, \tau_B \mathbf{v}_h)_{\partial\Omega} + \left(\mathbf{u}_h \cdot \mathbf{n}, \left(C_B \nu / h - \tau_B\right) \mathbf{v}_h \cdot \mathbf{n}\right)_{\partial\Omega}.
\end{aligned}
$$

Here, C_B, τ_B are user-chosen constants and we consider a choice at the end of this section. Let us modify the error analysis from Sect. 2, starting with a stability estimate. Please note that we obtain additional control of certain boundary terms.

Lemma 2. *Assume* $\mathbf{f} \in L^1(0,T; L^2(\Omega))$, $\mathbf{u}_0 \in [L^2(\Omega)]^d$, $C_B \geq \tau_B h/\nu > 0$ *and* $0 < \mu \leq \tau_B - 4C_I^2 \nu h^{-1}$. *Then we obtain*

$$\|\mathbf{u}_h\|_{L^\infty(0,t;L^2(\Omega))} \leq K(\mathbf{f},\mathbf{u}_0) := \|\mathbf{u}_0\|_0 + \|\mathbf{f}\|_{L^1(0,t;L^2(\Omega))}$$

$$\left|\left|\left|\mathbf{u}_h\right|\right|\right|_t^2 + \int_0^t \mu \|\mathbf{u}_h\|_{0,\partial\Omega}^2 + \frac{1}{2}\left\|\sqrt{|\mathbf{u}_h\cdot\mathbf{n}|}\mathbf{u}_h\right\|_{0,\Gamma_{in}}^2 dt \leq K^2(\mathbf{f},\mathbf{u}_0).$$

Proof. The new part in the proof is the treatment of the boundary terms. Therefore, we plug in $\mathbf{v}_h = \mathbf{u}_h$ and $q_h = p_h$ and obtain

$$\begin{aligned}
B_{\text{wall}}(\mathbf{u}_h, p_h; \mathbf{u}_h, p_h) &= -(2\nu\mathbf{D}\mathbf{u}_h\cdot\mathbf{n}, \mathbf{u}_h)_{\partial\Omega} + \left(p_h, \mathbf{u}_h\cdot\mathbf{n}\right)_{\partial\Omega} \\
&\quad - \frac{1}{2}((\mathbf{u}_h\cdot\mathbf{n})\mathbf{u}_h, \mathbf{u}_h)_{\partial\Omega} - (\mathbf{u}_h, 2\nu\mathbf{D}\mathbf{u}_h\cdot\mathbf{n})_{\partial\Omega} - \left(p_h, \mathbf{u}_h\cdot\mathbf{n}\right)_{\partial\Omega} \\
&\quad + \frac{1}{2}(\mathbf{u}_h, (\mathbf{u}_h\cdot\mathbf{n})\mathbf{u}_h)_{\Gamma_{\text{out}}} + (\mathbf{u}_h, \tau_B\mathbf{u}_h)_{\partial\Omega} + \left(\mathbf{u}_h\cdot\mathbf{n}, \left(C_B\nu/h - \tau_B\right)\mathbf{u}_h\cdot\mathbf{n}\right)_{\partial\Omega} \\
&\geq C_B\nu/h\,\|\mathbf{u}_h\cdot\mathbf{n}\|_{0,\partial\Omega}^2 + \tau_B\left(\|\mathbf{u}_h\cdot\tau_1\|_{0,\partial\Omega}^2 + \|\mathbf{u}_h\cdot\tau_2\|_{0,\partial\Omega}^2\right) \\
&\quad - \nu\,\|\mathbf{D}\mathbf{u}_h\|_0^2 - 4C_I^2\nu/h\,\|\mathbf{u}_h\|_{0,\partial\Omega}^2 + \frac{1}{2}\left\|\sqrt{|\mathbf{u}_h\cdot\mathbf{n}|}\mathbf{u}_h\right\|_{0,\Gamma_{\text{in}}}^2 \\
&\geq \left(\tau_B - 4\nu C_I^2/h\right)\|\mathbf{u}_h\|_{0,\partial\Omega}^2 + \frac{1}{2}\left\|\sqrt{|\mathbf{u}_h\cdot\mathbf{n}|}\mathbf{u}_h\right\|_{0,\Gamma_{\text{in}}}^2 - \nu\,\|\mathbf{D}\mathbf{u}_h\|_0^2,
\end{aligned}$$

where τ_1, τ_2 are the directions orthogonal to $\mathbf{n}$ and C_I is from a trace theorem. With the usual way to handle the other terms, e.g. in Ref. [7], the claim is proven.

Now we are in the position to state the following semidiscrete a-priori error estimate.

Theorem 2. *Assume $V_h = [\mathbf{Q}_k]^3$, $Q_h = \mathbf{Q}_{k-1}$, $k \geq 2$ and $\nu_T^K(\mathbf{u}_h) \in \mathcal{O}(h^2)$. Under the assumptions of Lemma 2 and sufficient smoothness assumptions on the data, see Ref. [7], it holds*

$$\left|\left|\left|\mathbf{u}_h - I_h\mathbf{u}\right|\right|\right|_t^2 + \int_0^t \mu\,\|\mathbf{u}_h - I_h\mathbf{u}\|_{0,\partial\Omega}^2 + \frac{1}{2}\left\|\sqrt{|\mathbf{u}_h\cdot\mathbf{n}|}\,(\mathbf{u}_h - I_h\mathbf{u})\right\|_{0,\Gamma_{in}}^2 dt \leq Ch^{2k}$$

for every $t \in (0,T]$ with $C = C(\nu, \nu_T, T, \mathbf{u})$.

Proof. Let us show the key steps. Please note that $B_{\text{wall}}(\mathbf{u}, p; \mathbf{v}_h, q_h) = 0$. The boundary terms can be treated like linear terms since $\mathbf{u}\cdot\mathbf{n}|_{\partial\Omega} = 0$. Consider $\mathbf{e}_u := \mathbf{u}_h - I_h\mathbf{u}$ and $\varepsilon_u := \mathbf{u} - I_h\mathbf{u}$ with an interpolation operator $I_h : H^1(\Omega) \to \widetilde{V}_h^{\text{div}}$ with

$$\widetilde{V}_h^{\text{div}} := \{\mathbf{v}_h \in V_h : (\nabla\cdot\mathbf{v}_h, q_h) = 0,\ (\mathbf{v}_h\cdot\mathbf{n}, q_h)_{\partial\Omega} = 0\ \forall q_h \in Q_h\}.$$

Such an operator can be constructed by minor modifications of the operator from Ref. [3]. With the exception of the boundary terms, everything can be estimated exactly as in the case of strong boundary conditions, see Ref. [7] formula (15). Hence, we focus on the boundary terms

$$B_{\text{wall}}(\mathbf{u}_h - \mathbf{u},\, p_h - p; \mathbf{e}_u, e_p) = B_{\text{wall}}(\mathbf{e}_u, e_p; \mathbf{e}_u, e_p) - B_{\text{wall}}(\varepsilon_u, \varepsilon_p; \mathbf{e}_u, e_p),$$

where the first term on the right hand side can be treated like in Lemma 2 to get

$$B_{\text{wall}}(\mathbf{e}_u, e_p; \mathbf{e}_u, e_p) \geq \left(\tau_B - 4\nu C_I^2/h\right) \|\mathbf{e}_u\|_{0,\partial\Omega}^2 + \left(C_B\nu/h - \tau_B\right) \|\mathbf{e}_u \cdot \mathbf{n}\|_{0,\partial\Omega}^2 + \frac{1}{2} \left\| \sqrt{|\mathbf{u}_h \cdot \mathbf{n}|}\mathbf{e}_u \right\|_{0,\Gamma_{\text{in}}}^2 - \nu \|\mathbf{D}\mathbf{e}_u\|_0^2 .$$

The structure of $B_{\text{wall}}(\varepsilon_u, \varepsilon_p; \mathbf{e}_u, e_p)$ is a combination of interpolation errors and terms for the left hand side. We estimate

$$\begin{aligned} B_{\text{wall}}(\varepsilon_u, \varepsilon_p; \mathbf{e}_u, e_p) \leq{}& \frac{\nu}{16} \|\mathbf{D}\mathbf{e}_u\|_0^2 + \frac{\mu}{2} \|\mathbf{e}_u\|_{0,\partial\Omega}^2 + \left(C_B\nu/h - \tau_B\right) \|\mathbf{e}_u \cdot \mathbf{n}\|_{0,\partial\Omega}^2 \\ &+ \frac{1}{4} \left\| \sqrt{|\mathbf{u}_h \cdot \mathbf{n}|}\mathbf{e}_u \right\|_{0,\Gamma_{\text{in}}}^2 + \frac{4\nu^2 C_I^2}{\mu h} \|\mathbf{D}\varepsilon_u\|_0^2 \\ &+ \left(16\nu C_I^2/h + \tau_B^2/\mu\right) \|\varepsilon_u\|_{0,\partial\Omega}^2 + \frac{1}{2} \left(C_B\nu/h - \tau_B\right) \|\varepsilon_u \cdot \mathbf{n}\|_{0,\partial\Omega}^2 \\ &+ \frac{1}{4} \left\| \sqrt{|\mathbf{u}_h \cdot \mathbf{n}|}\varepsilon_u \right\|_{0,\Gamma_{\text{in}}}^2 + \frac{1}{2} \left(C_B\nu/h - \tau_B\right)^{-1} \|\varepsilon_p\|_{0,\partial\Omega}^2 . \end{aligned}$$

To obtain this result we use $(e_p, (\mathbf{u} - I_h\mathbf{u}) \cdot \mathbf{n})_{\partial\Omega} = 0$, since we chose $I_h\mathbf{u} \in \widetilde{V}_h^{\text{div}}$. Together with the techniques in Ref. [7] and the stability result in Lemma 2, the claim is proven.

Let us discuss a choice of the parameters C_B and τ_B. The variant of weakly imposed boundary conditions from Ref. [1] is based on a wall function formulation, where $\tau_B = u_\tau^2/\|\mathbf{u}_{h,\tan}\|$ is determined to fulfill Spalding's law of the wall for a turbulent boundary layer with the wall-friction velocity $u_\tau^2 = \nu \frac{\partial <u_1>}{\partial y}|_{y=0}$ and the velocity vector $\mathbf{u}_{h,\tan}$ tangential to the wall. For the channel flow at $Re_\tau = 180$, one obtains $u_\tau \approx 0.0028$ and $\|\mathbf{u}_{h,\tan}\| \approx 0.043$, see Ref. [6]. For a better understanding of this parameter we refer to the theory on boundary layers. In the viscous sublayer it holds $1 = y^+/u^+ = y\tau_B/\nu$, where the '+' stands for the wall coordinates and y is the coordinate normal to the wall. This means that $\tau_B \sim \nu/h$ in the viscous sublayer and $\tau_B \geq C\nu/h$ away from layer, since $u^+ < y^+$ in these regions. From the analysis above we need the existence of $0 < \mu \leq \tau_B - 4C_I^2\nu/h$. A possible choice of τ_B is

$$\tau_B = \max\left(u_\tau^2/\|\mathbf{u}_{h,\tan}\|, 8C_I^2\nu/h\right).$$

For the remaining parameter, the analysis leads to $C_B > \tau_B h/\nu$.

5 Summary and Outlook

We applied a projection-based variational multiscale method to the numerical simulation of wall-bounded flows at moderate Reynolds numbers. A semidiscrete a-priori error estimate from [7] is extended to layer-adapted meshes of tensor-product type and to weak Dirichlet boundary conditions on isotropic meshes. Based on the error analysis, model parameters for the weak boundary treatment are derived.

For the channel flow at $Re_\tau = 180$ with layer-adapted meshes of tensor-product type no instability was obtained for meshes with moderately high aspect ratio. The numerical results of [1] for a channel flow on isotropic meshes validate the application of a weak treatment of Dirichlet boundary condition at even higher Re_τ.

For higher Reynolds numbers, a weak treatment of Dirichlet boundary conditions on isotropic meshes seems to be more advantageous as it mimics the wall of the law. Despite the potential influence of large aspect ratios, layer-adapted meshes will be more and more expensive with increasing Reynolds numbers.

Acknowledgements The research of Lars Röhe was supported by the German Research Foundation (DFG) through Research Training Group 1023.

References

1. Bazilevs, Y., Michler, C., Calo, V.M., Hughes, T.J.R, Isogeometric variational multiscale modeling of wall-bounded turbulent flows with weakly-enforced boundary conditions on unstretched meshes. *Comput. Meths. Appl. Mech. Engrg.* 199 (2010), 780–790.
2. Braack, M., Lube, G., Röhe, L., Divergence preserving interpolation on anisotropic quadrilateral meshes. *Comput. Methods Appl. Math.* 12 (2012) 2, 123–138.
3. Girault, V., Scott, L., A quasi-local interpolation operator preserving the discrete divergence. *Calcolo* 40 (2003), 1–19.
4. John, V., Kindl, A., Numerical studies of finite element variational multiscale methods for turbulent flow simulations. *Comput. Meths. Appl. Mech. Engrg.* 199 (2010) 13–16, 853–864.
5. Layton, W.J., A connection between subgrid scale eddy viscosity and mixed methods. *Appl. Math. Comput.* 133 (2002), 147–157.
6. Moser, Kim, Mansour, Direct numerical simulation of a turbulent channel flow up to $Re_\tau = 590$. *Phys. Fluids* 11 (1999) 4, 943–945.
7. Röhe, L., Lube, G., Analysis of variational multiscale method for large-eddy simulation and its application to homogeneous isotropic turbulence. *Comput. Meths. Appl. Mech. Engrg.* 199 (2010), 2331–2342.

On Higher-Order Space-Time Discretization of an Nonlinear Aeroelastic Problem with the Consideration of Large Displacements

P. Sváček

Abstract This paper focuses on the mathematical and numerical modelling of interaction of the two-dimensional incompressible fluid flow and a flexibly supported airfoil section with a control section. A simplified problem is considered: The flow is modelled by the system of Navier-Stokes equations and and the structure motion is described with the aid of nonlinear ordinary differential equations. The time-dependent computational domain is taken into account by the Arbitrary Lagrangian-Eulerian method. Higher order time discretization is considered within the stabilized finite element method. The application of the described method is shown.

1 Introduction

The problem of mutual interactions of fluid flow with elastic structure motion is important in many technical and scientific applications, e.g. [11]. Mostly, the simplified linearized problems are used, but in specific applications the solution of nonlinear coupled problems can be important, e.g. post-flutter behaviour, behaviour at the edge of instability of a structure, etc. The nonlinearities in the aeroelastic model can be of an aerodynamic nature or of the structural origin. In this paper we consider the nonlinearity involved in the stiffness matrices due to high vibration amplitudes and also structural nonlinearities in the stiffness matrix, see also [5].

This paper focuses on the extension of previously published analysis of flow induced vibrations of an airfoil with two degrees of freedom(DOF) or 3-DOF, see [5, 20]. Similar problem were studied in a number of scientific papers by different methods. Particularly, in [22] the motion of airfoil with 3-DOF was studied with

P. Sváček (✉)
Faculty of Mechanical Engineering, Department of Technical Mathematics, Czech Technical University in Prague, Karlovo nám. 13, Praha 2, 121 35, Czech Republic
e-mail: Petr.Svacek@fs.cvut.cz

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_63,

the aid of the Theodorsen's theory and the Wagner's function the piecewise linear equations of motion for the airfoil with control surface free play nonlinearity were established and the characteristics of limit cycle oscillations (LCO) were predicted by numerical simulation.

Analytical analysis and numerical simulations of the aeroelastic response of 3-DOF airfoil-flap system subjected to time-dependent loads in an incompressible subsonic flow were addressed in [18], where feedback control methodology in order to suppress the flutter instability was studied. The aeroelastic response and the active control of 3-DOF airfoil exposed to time-dependent external excitations in a subsonic compressible flow field were studied in [9]. Similarly, the later study in [10] was devoted to the aeroelastic response and control of 3-DOF flapped-wing system in an incompressible fluid flow and exposed to external pressure pulse. The goal was to suppress flutter instability and reduce the vibration level. The paper [7] presents an analysis of control strategies applied to a nonlinear 2-D aeroelastic 3-DOF wing-flap system operating in supersonic flow. The effectiveness in reducing the aeroelastic vibrations and in suppressing flutter is demonstrated. Structural nonlinearities can have significant effects on the aeroelastic responses even for small vibration amplitudes. Recently, [1] analyzed limit cycle oscillations (LCO) for a 2-DOF airfoil motion in plunge and pitch containing a hysteresis structural nonlinearity.

This paper focuses on the mathematical modelling and the numerical approximation of interactions of a simplified problem of two-dimensional flow and a flexibly supported airfoil section with control section. The flow is modelled with the aid of the incompressible Navier-Stokes equations and for the approximation the stabilized finite element method is used, cf. [2, 6]. For different stabilization approaches see also [8, 13, 14] or [21]. The structure motion is described with the aid of nonlinear ordinary differential equations. The time-dependent computational domain is taken into account by the Arbitrary Lagrangian-Eulerian method, cf. [12], or [20] for application in the aeroelastic application. The method is applied on a benchmark problem studied previously in [5]. In this paper the main attention is paid to application of higher order time discretizations within a stabilized finite element method for solution of an aeroelastic problem. The comparison of different time discretizations is presented.

2 Mathematical Model

In what follows we shall be interested in approximation of incompressible fluid flow in time-dependent computational domain $\Omega_t \subset R^2$, $t \in [0, T]$. In order to treat the time discretization the Arbitrary Lagrangian-Eulerian (ALE) method shall be used. We assume that there exists a smooth one-to-one transformation $\mathcal{A}_t$ (ALE mapping) of a reference computational domain Ω_0 onto Ω_t for $t \in (0, T)$. Further, by $\mathbf{w}_D(x, t)$ the *(ALE) domain velocity* and by $D^{\mathcal{A}}/Dt$ the *ALE derivative* shall be denoted. For more details about ALE method see, e.g., [12], or [5, 20]. The flow is

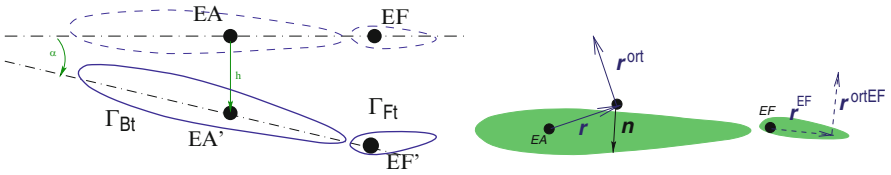


Fig. 1 Scheme of an airfoil with aileron in a deformed position

modelled by the incompressible Navier-Stokes equations

$$\begin{aligned} D^{\mathcal{A}}\mathbf{v}/Dt \; - \nu\triangle\mathbf{v} + ((\mathbf{v}-\mathbf{w}_D)\cdot\nabla)\mathbf{v} + \nabla p &= 0 \qquad \text{in } \Omega_t,\ t\in(0,T), \\ \operatorname{div}\mathbf{v} &= 0 \qquad \text{in } \Omega_t,\ t\in(0,T), \end{aligned} \tag{1}$$

equipped with initial and boundary conditions

$$\begin{aligned} &\text{a)} && \mathbf{v} = \mathbf{v}^0 && \text{in } \Omega_0, t=0, \\ &\text{b)} && \mathbf{v} = \mathbf{v}_D && \text{on } \Gamma_D, t\in(0,T), \\ &\text{c)} && \mathbf{v} = \mathbf{w}_D && \text{on } \Gamma_{Wt}, t\in(0,T) \\ &\text{d)} && -(p-p_{ref})\mathbf{n} + \tfrac{1}{2}(\mathbf{v}\cdot\mathbf{n})^{-}\mathbf{v} + \nu\tfrac{\partial\mathbf{v}}{\partial\mathbf{n}} = 0 && \text{on } \Gamma_O, t\in(0,T). \end{aligned} \tag{2}$$

Here, $\mathbf{v} = \mathbf{v}(x,t)$ denotes the velocity vector, $p = p(x,t)$ denotes the kinematic pressure (i.e. pressure divided by the constant fluid density ρ), ν denotes the kinematic viscosity, $\mathbf{n}$ denotes the unit outward normal vector to Ω_t, p_{ref} denotes a reference pressure value, and $\mathbf{v}_D$ is the Dirichlet boundary condition.

The motion of the structure is described by the system of linear ordinary differential equations for vertical displacement $h = h(t)$ (downwards positive), angle of rotation of the airfoil $\alpha = \alpha(t)$ of the whole airfoil around the elastic axis (EA) (clockwise positive), and the angle of rotation of the aileron $\beta = \beta(t)$ around the elastic axis of the aileron (EF) (clockwise positive), see Fig. 1. The equations of motion for small displacements read (cf. [11])

$$\begin{aligned} \left[m\ddot{h} + S_\alpha\ddot{\alpha} + S_\beta\ddot{\beta}\right] \quad &+d_{hh}\dot{h} && +k_{hh}h = -L(t), \\ S_\alpha\cos\alpha\ddot{h} + I_\alpha\ddot{\alpha} + \left(I_\beta + \widetilde{\Delta}S_\beta\right)\ddot{\beta} \quad &+d_{\alpha\alpha}\dot{\alpha} && +k_{\alpha\alpha}\alpha = M(t), \\ S_\beta\ddot{h} + (I_\beta + \widetilde{\Delta}S_\beta)\ddot{\alpha} + I_\beta\ddot{\beta} \quad &+d_{\beta\beta}\beta && +k_{\beta\beta}\beta + \hat{k}_{\beta\beta}\beta^3 = M_\beta(t). \end{aligned} \tag{3}$$

Here, k_{hh}, $k_{\alpha\alpha}$, $k_{\beta\beta}$ are the stiffness coefficients, $\hat{k}_{\beta\beta}$ is the nonlinear stiffness coefficient, m is the mass of the airfoil, S_α is the static moment of the airfoil around the elastic axis EA, I_α is the inertia moment of the airfoil around EA, S_β is the static moment of the aileron around the elastic axis of the control section EF, I_β is the inertia moment of the aileron around EF, $\widetilde{\Delta}$ is the distance of EF from EA, see Fig. 1, and $L(t)$, $M(t)$ and $M_\beta(t)$ are the aerodynamical forces. Under the consideration of large values of α and β (similarly as in [20] for 2-DOF) also the

geometrical nonlinearities need do be included in the model leading to the nonlinear "mass matrix" on the left-hand side of the system of equations (3), see [5].

The aerodynamical lift force L acting in the vertical direction, the torsional moment M and the aerodynamical moment M_β acting on the control section part are defined by

$$L = -l\int\limits_{\Gamma_{Wt}} \sum_{j=1}^{2}\tau_{2j}n_j\,dS,\quad M = l\int\limits_{\Gamma_{Wt}} \sum_{i,j=1}^{2}\tau_{ij}n_j r_i^{\mathrm{ort}}\,dS,\quad M_\beta = l\int\limits_{\Gamma_{Ft}} \sum_{i,j=1}^{2}\tau_{ij}n_j r_i^{\mathrm{ortEF}}\,dS, \tag{4}$$

where $\tau_{ij} = \rho\left[-p\delta_{ij} + \nu\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right)\right]$, $r_1^{\mathrm{ort}} = -(x_2 - x_{\mathrm{EO2}})$, $r_2^{\mathrm{ort}} = x_1 - x_{\mathrm{EO1}}$, $r_1^{\mathrm{ortEF}} = -(x_2 - x_{\mathrm{EF2}})$, $r_2^{\mathrm{ortEF}} = x_1 - x_{\mathrm{EF1}}$. The position of the elastic axis is given by $x_{\mathrm{EO}} = (x_{\mathrm{EO1}}, x_{\mathrm{EO2}})$ and by $x_{\mathrm{EF}} = (x_{\mathrm{EF1}}, x_{\mathrm{EF2}})$ the position of the elastic axis of the control section is denoted, see Fig. 1. By l the depth of the considered wing section is denoted.

The structure and flow models are mutually coupled due to kinematic boundary condition on the airfoil surface ($\mathbf{v} = \mathbf{w}_D$). Further, the aerodynamical lift force L, the torsional moment M of the airfoil (A) and the torsional moment M_β of the aileron (F) are computed from flow quantities ($\mathbf{v}$, p) at any time t.

3 Numerical Approximation

Time discretization. In order to discretize the problem (1) the uniform partition of the time interval $(0, T)$ with the constant time step Δt is considered and by $\mathbf{v}^n$ the approximation of the time dependent function at time level $t_n = n\Delta t$ is denoted. Similarly, p^n denotes the approximation of the pressure at time instant t_n. The ALE derivative is then approximated with the aid of higher order backward difference formulae (BDFq), see [15]. The application of BDF2 was described previously in [5]. Here, the third (BDF3) and fourth(BDF4) order formulae are considered defined by

$$\frac{D^{\mathscr{A}}\mathbf{v}}{Dt}(x,t) \approx \frac{11\mathbf{v}^{n+1} - 18\widehat{\mathbf{v}^n} + 9\widehat{\mathbf{v}^{n-1}} - 2\widehat{\mathbf{v}^{n-2}}}{6\Delta t},$$

and

$$\frac{D^{\mathscr{A}}\mathbf{v}}{Dt}(x,t) \approx \frac{25\mathbf{v}^{n+1} - 48\widehat{\mathbf{v}^n} + 36\widehat{\mathbf{v}^{n-1}} - 16\widehat{\mathbf{v}^{n-2}} + 3\widehat{\mathbf{v}^{n-3}}}{12\Delta t},$$

where by $\widehat{\mathbf{v}^{n-j}}$ the flow velocity at time instant t_{n-j} transformed from $\Omega_{t_{n-j}}$ onto $\Omega_{t_{n+1}}$ is denoted. Generally, we write the BDFq formula as

$$\frac{D^{\mathscr{A}}\mathbf{v}}{Dt}(x,t) \approx \frac{\beta_{-1}\mathbf{v}^{n+1}}{\Delta t} - \frac{1}{\Delta t}\sum_{j=0}^{q-1}\beta_j\widehat{\mathbf{v}^{n-j}}. \tag{5}$$

Similarly, the system of ODEs (3) is transformed to first order system $\dot{Y} = KY + f_{non}(Y)$ and time discretized with the same BDFq formula. This leads to a system of non-linear equations, which is is then solved by fixed-point iterations (for the solution only few steps are necessary, for the given numerical example 2–4 steps were needed).

Spatial discretization. Further, system of Eq. (1) is formulated weakly and the solution is sought in a couple of finite element spaces. Let us mention that the finite element spaces should satisfy the *Babuška–Brezzi (BB) condition* (see, e.g., [16, 17]). In practical computations we assume that the domain $\Omega = \Omega_{t_{n+1}}$ is a polygonal approximation of the region occupied by the fluid at time t_{n+1} and the finite element spaces are defined over a triangulation $\mathscr{T}_\Delta$ of the domain Ω as piecewise polynomial functions. In order to capture thin boundary layer and other phenomena, (a priori or a posteriori) anisotropic mesh refinement needs to be used, cf. [4]. In our computations, the well-known Taylor-Hood P_2/P_1 conforming elements are used for the velocity/pressure approximation. This means that p_Δ is a linear function and $\mathbf{v}_\Delta$ is a quadratic vector-valued function on each element $K \in \mathscr{T}_\Delta$. The weak formulation of the time discretized problem (1) is then represented by the Galerkin terms (at time level t_{n+1} on $\Omega = \Omega_{t_{n+1}}$)

$$\mathbf{a}(U^*; U, V) = \frac{\beta_{-1}}{\Delta t}(\mathbf{v}, \mathbf{z})_\Omega + \nu(\nabla \mathbf{v}, \nabla \mathbf{z})_\Omega + (\overline{\mathbf{w}} \cdot \nabla \mathbf{v}, \mathbf{z})_\Omega - \big(p, \nabla \cdot \mathbf{z}\big)_\Omega + \big(\nabla \cdot \mathbf{v}, q\big)_\Omega,$$

$$f(V) = \frac{1}{\Delta t} \sum_{j=0}^{q-1} \left(\beta_j \widehat{\mathbf{v}^{n-j}}, \mathbf{z}\right)_\Omega,$$

where $U = (\mathbf{v}, p)$, $V = (\mathbf{z}, q)$, $U^* = (\mathbf{v}^*, p^*)$, and $\overline{\mathbf{w}}$ stands for the transport velocity, i.e. $\overline{\mathbf{w}} = \mathbf{v}^* - \mathbf{w}^{n+1}$, $(\cdot, \cdot)_\omega$ denotes the scalar product in $L^2(\omega)$. Further, in order to obtain stable solution even for very large values of Reynolds numbers the SUPG/PSPG stabilization of FEM shall be used. To this end the additional stabilizing terms are defined by

$$\mathscr{L}_\Delta(U^*, U, V) = \sum_{K \in \mathscr{T}_\Delta} \delta_K \Big(\frac{\beta_{-1}}{\Delta t}\mathbf{v} - \nu \triangle \mathbf{v} + (\overline{\mathbf{w}} \cdot \nabla)\,\mathbf{v} + \nabla p, (\overline{\mathbf{w}} \cdot \nabla)\mathbf{v}\Big)_K,$$

$$\mathscr{F}_\Delta(V) = \sum_{K \in \mathscr{T}_\Delta} \delta_K \Big(\frac{1}{\Delta t}\Big(\sum_{j=0}^{q-1} \beta_j \widehat{\mathbf{v}^{n-j}}, (\overline{\mathbf{w}} \cdot \nabla)\mathbf{v}\Big)_K, \tag{6}$$

the additional grad-div stabilization and the stabilizing parameters reads

$$\mathscr{P}_\Delta(U, V) = \sum_{K \in \mathscr{T}_\Delta} \tau_K (\nabla \cdot \mathbf{v}, \nabla \cdot \mathbf{z})_K, \qquad \delta_K = \delta^* {h_K}^2, \qquad \tau_K = 1. \tag{7}$$

where the choice of the parameters δ_K and τ_K is carried out according to [19] on the basis of the local element length h_K.

The *stabilized discrete problem* reads: Find $U_\Delta = (\mathbf{v}_\Delta, p_\Delta) \in W_\Delta \times Q_\Delta$ such that $\mathbf{z}_\Delta$ satisfies approximately conditions (2b), (2c) and

$$a(U_\Delta, U_\Delta, V_\Delta) + \mathscr{L}_\Delta(U_\Delta, U_\Delta, V_\Delta) + \mathscr{P}_\Delta(U_\Delta, V_\Delta) = f(V_\Delta) + \mathscr{F}_\Delta(V_\Delta)$$
$$\text{for all } V_\Delta = (\mathbf{z}_\Delta, q_\Delta) \in X_\Delta \times Q_\Delta. \tag{8}$$

The stabilized discrete problem is then linearized by Oseen iterations, and each linearization is solved with the aid of direct solver, cf. [3]. Following our experience only few Oseen iterations are needed, for the numerical example bellow the number of Oseen iterations were 3-6 and the convergence to the given precision was obtained.

4 Numerical Results

The developed technique was used for numerical simulation of an aeroelastic problem with three degrees of freedom, see [5]. The following choice of parameters was employed $k_{hh} = 105.1$ N/m, $k_{\alpha\alpha} = 3.7$ N m/rad, $k_{\beta\beta} = 0.2$ N m/rad, $m = 0.0866\,kg$, $S_\alpha = -0.0007796$ kg/m and $I_\alpha = 0.000487$ kg m^2. $S_\beta = 0$ kg/m and $I_\beta = 3.411 \times 10^{-5}$ kg m^2. The elastic axis (EA) is located at 40 % of the airfoil and the elastic axis of the flap section (EF) is located at 80 % of the airfoil. The depth of the considered section is 7.9 cm. The linear theory predicts the critical velocity $U_{crit} = 11.3$ m/s.

The numerical results are shown in Figs. 2–5. The comparison of behaviour of BDFq for the case of problem with 2DOF is shown in Figs. 2–3. For subcritical velocities and small time step the results of BDF2 and BDF3 are almost identical Fig. 2. For velocity closer to the critical velocity nevertheless the higher order formulae BDF3/BDF4 leeds to more precise results, see Fig. 3. The comparison of BDF2/BDF3 results for the case of 3DOF is shown in Fig. 4. The comparison shows, that the BDF3 formula is easy to be used, and still can provide precise numerical solution without loosing stability of the method. On the other hand, the application of BDF4 seems to be questionable as the formula is only conditionally stable provided that the time step is small enough. In such cases the numerical results obtained by BDF4 are almost same as for BDF3.

The aeroelastic response in terms of α, h and β in dependence on time can be seen in Fig. 5 obtained by BDF3 formula. The results shows the weakly damped vibrations of the airfoil for the far field velocity being very close to the critical velocity. For this case, Fig. 6 shows the distribution of the flow velocity magnitude at four different time instants during one cycle.

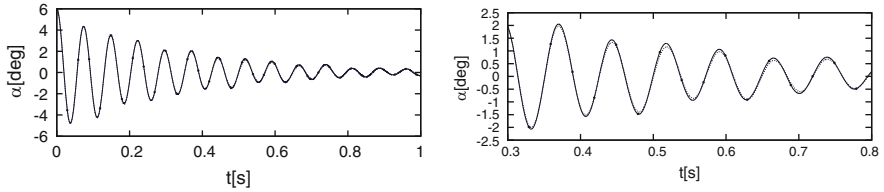


Fig. 2 Comparison of the numerical results for the aeroelastic problem with 2 DOF in terms of the aeroelastic response $\alpha(t)$ obtained by BDF2 (*solid line*) and BDF3 (*solid line* with symbols) formulae for $U_\infty = 10$ m/s

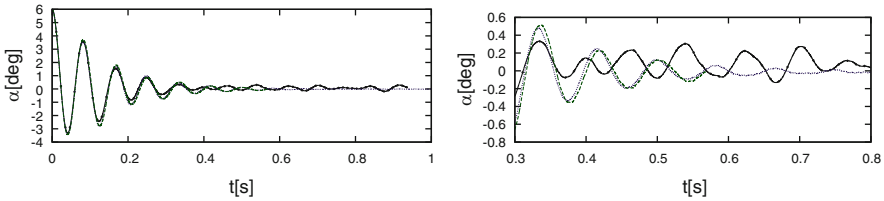


Fig. 3 Comparison of the numerical results for the aeroelastic problem with 2 DOF in terms of the aeroelastic response $\alpha(t)$ obtained by BDF2 (*solid line*), BDF3 (*solid line* with symbols) and BDF4 (*dashed line*) formulae for $U_\infty = 25$ m/s

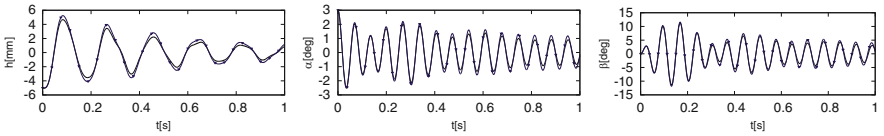


Fig. 4 Comparison of the numerical results for the aeroelastic problem with 3 DOF in terms of the aeroelastic response $\alpha(t)$ obtained by BDF2 (*solid line*) and BDF3 (*solid line* with symbols) formulae for $U_\infty = 8$ m/s

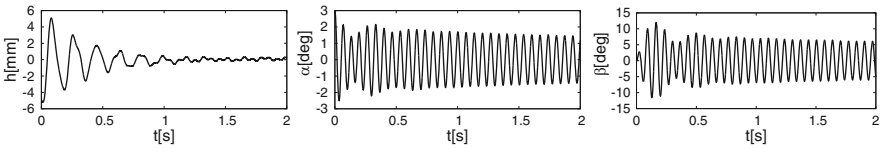


Fig. 5 The aeroelastic response h, α, β for $U_\infty = 11$ m/s

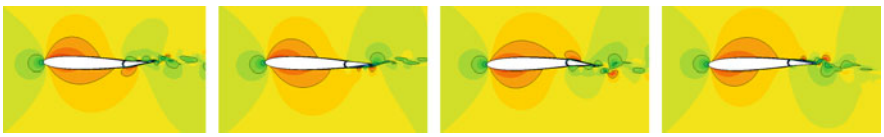


Fig. 6 Flow velocity magnitude distribution during four time instants for far field velocity $U_\infty = 11$ m/s

Acknowledgements This research was supported under grants No. 201/08/0012 and No. P101/11/0207 of the Grant Agency of the Czech Republic.

References

1. Chung, K., He, Y., Lee, B.: Bifurcation analysis of a two-degree-of-freedom aeroelastic system with hysteresis structural nonlinearity by a perturbation-incremental method. Journal of Sound and Vibration **320**(1–2), 163–183 (2009)
2. Codina, R.: Stabilization of incompressibility and convection through orthogonal sub-scales in finite element methods. Computational Method in Applied Mechanical Engineering **190**, 1579–1599 (2000)
3. Davis, T.A., Duff, I.S.: A combined unifrontal/multifrontal method for unsymmetric sparse matrices. ACM Transactions on Mathematical Software **25**, 1–19 (1999)
4. Dolejší, V.: Anisotropic mesh adaptation technique for viscous flow simulation. East-West Journal of Numerical Mathematics **9**, 1–24 (2001)
5. Feistauer, M., Horáček, J., Růžička, M., Sváček, P.: Numerical analysis of flow-induced nonlinear vibrations of an airfoil with three degrees of freedom. Computers & Fluids **49**(1), 110–127 (2011).
6. Gelhard, T., Lube, G., Olshanskii, M.A., Starcke, J.H.: Stabilized finite element schemes with LBB-stable elements for incompressible flows. Journal of Computational and Applied Mathematics **177**, 243–267 (2005)
7. Kim, K., Lee, B., Na, S., Marzocca, P., Milanese, A.: Comparative analysis of control performances applied to a 3-dofs nonlinear supersonic lifting surface. In: AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference, 1724 (2008)
8. Lube, G., Rapin, G.: Residual-based stabilized higher-order fem for advection-dominated problems. Computer Methods in Applied Mechanics and Engineering **195**(33–36), 4124–4138 (2006).
9. Na, S., Jeong, I.J., Librescu, L., Marzocca, P.: Aeroelastic response and active control of an airfoil in subsonic compressible flow. Collection of Technical Papers - AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference **4**, 2406–2420 (2005)
10. Na, S., Librescu, L., Kim, M.H., Jeong, I.J., Marzocca, P.: Robust aeroelastic control of flapped wing systems using a sliding mode observer. Aerospace Science and Technology **10**(2), 120–126 (2006)
11. Naudasher, E., Rockwell, D.: Flow-Induced Vibrations. A.A. Balkema, Rotterdam (1994)
12. Nomura, T., Hughes, T.J.R.: An arbitrary Lagrangian-Eulerian finite element method for interaction of fluid and a rigid body. Computer Methods in Applied Mechanics and Engineering **95**, 115–138 (1992)
13. Olshanskii, M., Lube, G., Heister, T., Löwe, J.: Grad–div stabilization and subgrid pressure models for the incompressible navier–stokes equations. Computer Methods in Applied Mechanics and Engineering **198**(49–52), 3975–3988 (2009).
14. Onate, E., Valls, A., García, J.: Modeling incompressible flows at low and high reynolds numbers via a finite calculus–finite element approach. Journal of Computational Physics **224**(1), 332–351 (2007).
15. Owen, H., Codina, R.: A third-order velocity correction scheme obtained at the discrete level. International Journal for Numerical Methods in Fluids (2011).
16. Raviart, P.A., Girault, V.: Finite Element Methods for the Navier-Stokes Equations. Springer, Berlin (1986)
17. Sani, R.L., Gresho, P.M.: Incompressible Flow and the Finite Element Method. Wiley, Chichester (2000)

18. Shim, J.H., Na, S., Chung, C.H.: Aeroelastic response of an airfoil-flap system exposed to time-dependent disturbances. KSME International Journal **18**(4), 560–572 (2004)
19. Sváček, P., Feistauer, M.: Application of a Stabilized FEM to Problems of Aeroelasticity. In: Numerical Mathematics and Advanced Application, pp. 796–805. Springer, Berlin (2004)
20. Sváček, P., Feistauer, M., Horáček, J.: Numerical simulation of flow induced airfoil vibrations with large amplitudes. Journal of Fluids and Structure **23**(3), 391–411 (2007)
21. Tezduyar, T., Osawa, Y.: Finite element stabilization parameters computed from element matrices and vectors. Computer Methods in Applied Mechanics and Engineering **190**(3), 411–430 (2000).
22. Zhao, Y.H., Hu, H.Y.: Aeroelastic analysis of a non-linear airfoil based on unsteady vortex lattice model. Journal of Sound and Vibration **276**(3-5), 491–510 (2004).

On the Construction of Analytic Solutions to a Visco–Elasticity Model for Soft Tissues

F.J. Vermolen

Abstract We consider a simple model for visco–elasticity, that is commonly applied to simulate dermal wound healing. First the problem is formulated, then, convergence to a steady–state equilibrium solution is demonstrated. Subsequently, we construct analytic solutions based on Green's Functions for one-dimensional sample problems. These solutions enable us to look at the convergence behavior towards equilibrium solutions. We also give some conditions for monotonic convergence.

1 Introduction

Visco–elastic materials exist in various applications and natural environments such as polymers or soft tissues. The present paper is mainly devoted to the application to soft tissues where one commonly uses the Kelvin–Voigt model for visco–elasticity. Some studies exist where visco–elasticity models are analyzed in terms of existence and uniqueness of solutions, see for instance studies by Eck and Jarusek [1] and Gilbert et al.[2]. The Kelvin–Voigt model for visco–elasticity is based on a linear elasticity model, via Hooke's Law, combined with a first–order time–derivative of spatial derivatives. This model is commonly applied in contraction models for dermal wound healing, such as [3] to mention an example. Hence, many wound healing models as well as models for tumor growth in soft tissues will involve the visco–elastic equations. In all these studies, the visco–elastic equations are solved by the use of discretization techniques as finite–element or finite–difference methods combined with conservation laws for cellular densities and chemical agents like growth factors or oxygen. However, a qualitative analysis

F.J. Vermolen (✉)
Delft Institute of Applied Mathematics, Delft University of Technology, Mekelweg 4, 2628 CD, Delft, The Netherlands
e-mail: F.J.Vermolen@tudelft.nl

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_64,

that consists of analytic solutions as well as convergence to steady–state equilibria is lacking, as far as we know. Therefore, the purpose of this paper is to draw some attention to the visco–elastic model in terms of presenting the partial differential equations, as well as the earlier mentioned mathematical properties. The analysis is useful for the understanding of the nature of solutions and can be used as a validation tool of numerical solutions. This paper contains some of the analysis that was used to validate the finite–element results in [3]. In this sense, this manuscript is complementary to [3].

In this paper, we will first pose the visco–elastic model, then, we will review some of the stability properties that hold in general. This is followed by the construction of analytic expressions in terms of Green's Functions. Finally we discuss the solutions and properties, and summarize with some conclusions.

2 The Visco–Elastic Model

In this section, we present the visco–elastic model in terms of a mechanical balance. We consider a bounded domain of computation $\Omega \subset \mathbb{R}^2$ with its boundary and closure, respectively, denoted by $\partial\Omega$ and $\overline{\Omega}$. Let ρ, $\underline{\underline{\sigma}}$, and $\underline{f}(x, y)$, respectively, be the local mass density, stress tensor and a body–force (such as gravity, or cellular traction forces in many soft tissue applications), then a mechanical balance gives

$$\rho \underline{u}_{tt} - \nabla \cdot \underline{\underline{\sigma}} = \underline{f}(x, y), \qquad (t, (x, y)) \in \mathbb{R}^+ \times \Omega, \tag{1}$$

where $\underline{u} = \underline{u}(t, x, y) = [u(t, x, y)\ v(t, x, y)]^T$ denotes the vector of displacement at position (x, y) and at time t with components $u(t, x, y)$ and $v(t, x, y)$, which is to be solved from the above equation. The stress tensor consists of both an elastic and a viscous part. Since, we are applying this model predominantly to soft tissue problems, we use the following set of initial and boundary conditions:

$$\begin{cases} \underline{u}(t, x, y) = \underline{0}, & \text{on } \partial\Omega_1, \\ \underline{\underline{\sigma}} \cdot \underline{n} = \underline{0}, & \text{on } \partial\Omega_2, \\ \underline{u}(0, x, y) = \underline{u}_0(x, y),\ \underline{u}_t(0, x, y) = \underline{q}_0(x, y), & \text{in } \Omega. \end{cases} \tag{2}$$

The analysis to be presented can also be carried out analogously if the arrangement of boundary conditions is altered. The stress tensor consists of a viscous, $\underline{\underline{\sigma}}_V$ and an elastic part, $\underline{\underline{\sigma}}_E$

$$\underline{\underline{\sigma}} = \underline{\underline{\sigma}}_V + \underline{\underline{\sigma}}_E, \tag{3}$$

where the viscous and elastic parts are given by

$$\begin{aligned} \underline{\underline{\sigma}}_V &= \mu_1 \underline{\underline{\varepsilon}}_t + \mu_2 (\nabla \cdot \underline{u}_t) \underline{\underline{I}}, \\ \underline{\underline{\sigma}}_E &= 2\mu \underline{\underline{\varepsilon}} + \lambda (\nabla \cdot \underline{u}) \underline{\underline{I}}, \end{aligned} \tag{4}$$

where μ_1, μ_2 are the kinematic and dynamic viscosity. The strain–tensor is denoted by $\underline{\underline{\varepsilon}}$. The above relation reflects linear elasticity, which is applicable as long as the strains are not too large. Further, μ and λ are the Lamé constants, which are related to the Young's Modulus, E, and Poisson Ratio, ν, by

$$\lambda = \frac{\nu E}{(1+\nu)(1-2\nu)}, \qquad \mu = \frac{E}{2(1+\nu)}. \tag{5}$$

The strain–tensor depends on the displacement by

$$\underline{\underline{\varepsilon}} = \frac{1}{2}\left(\nabla \underline{u} + (\nabla \underline{u})^T\right). \tag{6}$$

3 Stability of the Steady–State Equilibrium

Existence and uniqueness questions were dealt with by Eck and Jarusek [1] for $\rho = 0$ and more recently by Gilbert et al. [2], where $\rho \geq 0$ was considered for $\underline{u} \in L^\infty([0,T]; H^1(\Omega))$ using a Galerkin decomposition and a convergence argument, where we define the Bochner–space $L^\infty(I;X)$ by $L^\infty(I;X) := \{u \mid \operatorname*{ess\,sup}_{t \in I} ||u||_X < \infty\}$, where X is a Banach space, see page 285 in [4]. We demonstrate stability of the steady–state equilibrium. We consider the case $\rho = 0$ in Eq. (1), which physically corresponds to neglecting inertia. First, we consider the steady–state version of Eq. (1), which is given by

$$-\nabla \cdot \underline{\underline{\sigma_E}}(\underline{u}_E) = \underline{f}(x,y), \tag{7}$$

with boundary conditions

$$\begin{cases} \underline{u}_E(x,y) = \underline{0}, \text{ on } \partial\Omega_1, \\ \underline{\underline{\sigma}}_E \cdot \underline{n} = \underline{0}, \quad \text{on } \partial\Omega_2. \end{cases} \tag{8}$$

To this extent, we introduce the matrix inner product for two real–valued $n \times n$-matrices, given by

$$A : B := \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} b_{ij}, \tag{9}$$

where a_{ij} and b_{ij} represent the entries of the matrices A and B, respectively. Further, we will use Korn's Second Inequality (coercivity), which is given by Brenner and Scott [5, 6]:

Theorem 1. *Let $meas(\partial\Omega_1) > 0$ (a one-dimensional measure), then there exists a positive constant C such that*

$$\int_{\Omega} \underline{\underline{\varepsilon}}(\underline{u}) : \underline{\underline{\varepsilon}}(\underline{u}) d\Omega \geq C ||\underline{u}||^2_{H^1\Omega}, \tag{10}$$

for all $\underline{u} \in \Sigma := \{\underline{u} \in \underline{H}^1(\Omega) | \underline{u} = \underline{0}$ on $\partial\Omega_1\}$.

Using Korn's Second Inequality, Poincaré's Inequality, and the Lax–Milgram Lemma, existence and uniqueness of a solution in Σ to the weak form

$$\underline{u}_E \in \Sigma \text{ such that } \int_{\Omega} \underline{\underline{\sigma}}_E(\underline{u}_E) : \underline{\underline{\varepsilon}}(\underline{\phi}) d\Omega = \int_{\Omega} \underline{\phi} \cdot \underline{f}(x, y) d\Omega, \quad \forall \underline{\phi} \in \Sigma, \tag{11}$$

is demonstrated in a straightforward manner. After integration by parts, substitution of the boundary conditions, and some elementary algebra, we obtain the following weak form of Eq. (1) with initial and boundary conditions (2):

$$\begin{cases} \underline{u} \in L^2([0, T]; \Sigma) \text{ such that } b(\underline{u}_t, \underline{\phi}) + a(\underline{u}, \underline{\phi}) = (\underline{f}, \underline{\phi}), \text{ with } T > 0, \\ \text{for all } \underline{\phi} \in \Sigma, \text{ with } \underline{u}(0, x, y) = \underline{u}_0(x, y), \end{cases} \tag{12}$$

where the Bochner Space $L^p(I; X)$ (see also [4], p. 285), inner product $(.,.)$ and bilinear forms $a(.,.)$ and $b(.,.)$ are defined by

$$L^p(I; X) = \{u \mid \left(\int_I ||u(t,.)||^p_X dt\right)^{\frac{1}{p}} < \infty, \ (\underline{f}, \underline{\phi}) := \int_{\Omega} \underline{f} \cdot \underline{\phi} d\Omega,$$
$$a(\underline{u}, \underline{\phi}) := \int_{\Omega} \underline{\underline{\sigma}}(\underline{u}) : \underline{\underline{\varepsilon}}(\underline{\phi}) d\Omega, \qquad b(\underline{u}, \underline{\phi}) := \int_{\Omega} \mu_1 \underline{\underline{\varepsilon}}(\underline{u}) : \underline{\underline{\varepsilon}}(\underline{\phi}) + \mu_2 (\nabla \cdot \underline{u})(\nabla \cdot \underline{\phi}) d\Omega.$$

Next, we formulate the stability result for $\rho = 0$:

Theorem 2. *Let $\underline{u}$ be the solution to PDE (1), with $\rho = 0$, and with initial and boundary conditions from equation (2), and let $\underline{u}_E$ be the solution to PDE (7) with boundary conditions from equation (8). Then $\lim_{t\to\infty} \underline{u}(t, x, y) = \underline{u}_E(x, y)$ a.e. in Ω.*

Proof. Let $\underline{w} := \underline{u} - \underline{u}_E$. Then, $\underline{w}$ satisfies

$$\begin{cases} \underline{w} \in L^2([0, T]; \Sigma) : b(\underline{w}_t, \underline{\phi}) + a(\underline{w}, \underline{\phi}) = 0, \\ \text{subject to} \qquad\qquad \underline{w}(0, x, y) = \underline{u}_0 - \underline{u}_E. \end{cases} \tag{13}$$

Take $\underline{\phi} = \underline{w}$ at fixed time t, then, symmetry of $b(.,.)$ gives

$$\frac{1}{2} \frac{d}{dt} b(\underline{w}, \underline{w}) = -a(\underline{w}, \underline{w}). \tag{14}$$

Since $a(.,.)$ and $b(.,.)$ can be written as

$$\begin{aligned} a(\underline{w}, \underline{\phi}) &= \int_\Omega 2\mu \underline{\underline{\varepsilon}}(\underline{w}) : \underline{\underline{\varepsilon}}(\underline{\phi}) + \lambda(\nabla \cdot \underline{w})(\nabla \cdot \underline{\phi}) d\Omega, \\ b(\underline{w}, \underline{\phi}) &= \int_\Omega \mu_1 \underline{\underline{\varepsilon}}(\underline{w}) : \underline{\underline{\varepsilon}}(\underline{\phi}) + \mu_2(\nabla \cdot \underline{w})(\nabla \cdot \underline{\phi}) d\Omega, \end{aligned} \tag{15}$$

we can find $M > 0$ such that $a(\underline{\phi}, \underline{\phi}) \geq M b(\underline{\phi}, \underline{\phi})$, by imposing $M(2\mu + \lambda) \geq \mu_1 + \mu_2$. Hence we get

$$\frac{d}{dt} b(\underline{w}, \underline{w}) + 2M b(\underline{w}, \underline{w}) \leq 0. \tag{16}$$

By application of Grönwall's Lemma, we get

$$0 \leq b(\underline{w}, \underline{w}) \leq e^{-2Mt} b(\underline{u}_0 - \underline{u}_E, \underline{u}_0 - \underline{u}_E). \tag{17}$$

The left inequality follows from the fact that it is easy to see that $b(\underline{\phi}, \underline{\phi}) \geq \mu_1 \int_\Omega \underline{\underline{\varepsilon}} : \underline{\underline{\varepsilon}} d\Omega$, and hence by application of Korn's Second Inequality (see Theorem 1), it follows that the bilinear form $b(.,.)$ is coercive. This implies that $\lim_{t \to \infty} b(\underline{w}, \underline{w}) = 0$. Hence, combined with Lax–Friedrich's (also known as Poincaré's) Inequality and $\underline{w} = \underline{0}$ on $\partial\Omega$, we obtain $\underline{u} \to \underline{u}_E$ a.e. in Ω as $t \to \infty$. □

From this theorem, solutions to the visco–elastic equation always converge to the equilibrium steady–state. Further, the above argument is also easily applied to obtain uniqueness of solutions. We also note that it is straightforward to extend and to prove Theorem 2 for nonhomogeneous boundary conditions.

4 Analytic Solutions to the One-Dimensional Visco–Elasticity Model

We construct closed–form expressions for the one-dimensional visco–elastic equations where the following one-dimensional problem is used with $L > 0$:

$$\begin{cases} \rho u_{tt} - (\mu_1 + \mu_2) u_{xxt} - (\lambda + 2\mu) u_{xx} + \gamma u = f(x), & \text{for } (t, x) \in \mathbb{R}^+ \times (0, L), \\ u(t, 0) = 0, \quad u(t, L) = 0, & \text{for } t > 0, \\ u(0, x) = u_0(x), \quad u_t(0, x) = q_0(x), & \text{for } 0 < x < L. \end{cases} \tag{18}$$

We will also briefly state some results for different boundary conditions and remark that a travelling wave analysis in an unbounded domain can be found in [7] and in references therein. The procedures with the various boundary conditions are similar to the ones we will apply for the solution of the problem in the above equations. We construct a solution by considering convergence to the steady–state solution $u^E(x)$, from

$$\begin{cases} -(\lambda + 2\mu) u^E_{xx} + \gamma u^E = f(x), \text{ for } 0 < x < L, \\ u^E(0) = 0, \quad u^E(L) = 0. \end{cases} \tag{19}$$

The formal solution to the above steady–state problem can be obtained by means of elementary techniques. We illustrate this with the following example, which reflects various cases in dermal contraction models.

Example 1. We consider the following right–hand side function

$$f(x) = \begin{cases} 0, & 0 < x < R, \\ P, & R < x < R + \delta, \\ 0, & R + \delta < x < 1, \end{cases} \tag{20}$$

with $P > 0, 0 < R < R + \delta < L$, and denote $D = \lambda + 2\mu$. Then, for this problem, following [8], we have for $u^E(x)$:

$$u^E(x) = \begin{cases} a_1 e^{\sqrt{\frac{\gamma}{D}}x} + a_2 e^{-\sqrt{\frac{\gamma}{D}}x}, & 0 < x < R, \\ a_3 e^{\sqrt{\frac{\gamma}{D}}x} + a_4 e^{-\sqrt{\frac{\gamma}{D}}x} + PD, & R < x < R + \delta, \\ a_5 e^{\sqrt{\frac{\gamma}{D}}x} + a_6 e^{-\sqrt{\frac{\gamma}{D}}x}, & R + \delta < x < L. \end{cases}$$

The boundary conditions and continuity of u^E and $\frac{du^E}{dx}$ define the values for $a_1, \ldots, a_6$ uniquely.

$$a_1 = \frac{PD\left(e^{\sqrt{\frac{\gamma}{D}}(2L-R-\delta)} + e^{\sqrt{\frac{\gamma}{D}}R} - e^{\sqrt{\frac{\gamma}{D}}(R+\delta)} - e^{\sqrt{\frac{\gamma}{D}}(2L-R)}\right)}{2\gamma\left(e^{-\sqrt{\frac{\gamma}{D}}(R+\delta)} - e^{\sqrt{\frac{\gamma}{D}}(2L-R-\delta)}\right)} e^{-\sqrt{\frac{\gamma}{D}}(R+\delta)},$$

$$a_3 = \frac{PD\left(-e^{\sqrt{\frac{\gamma}{D}}(2L-R-\delta)} - e^{\sqrt{\frac{\gamma}{D}}R} + e^{\sqrt{\frac{\gamma}{D}}(R+\delta)} + e^{-\sqrt{\frac{\gamma}{D}}R}\right)}{2\gamma\left(-e^{-\sqrt{\frac{\gamma}{D}}(R+\delta)} + e^{\sqrt{\frac{\gamma}{D}}(2L-R-\delta)}\right)} e^{-\sqrt{\frac{\gamma}{D}}(R+\delta)},$$

$$a_4 = \frac{PD\left(-e^{\sqrt{\frac{\gamma}{D}}(2L-R-\delta)} + e^{\sqrt{\frac{\gamma}{D}}(2L-R)} + e^{\sqrt{\frac{\gamma}{D}}(R+\delta)} - e^{-\sqrt{\frac{\gamma}{D}}(2L+R)}\right)}{2\gamma\left(-e^{-\sqrt{\frac{\gamma}{D}}(R+\delta)} + e^{\sqrt{\frac{\gamma}{D}}(2L-R-\delta)}\right)} e^{-\sqrt{\frac{\gamma}{D}}(R+\delta)},$$

$$a_5 = \frac{PD\left(e^{\sqrt{\frac{\gamma}{D}}R} - e^{-\sqrt{\frac{\gamma}{D}}R} - e^{\sqrt{\frac{\gamma}{D}}(R+\delta)} + e^{-\sqrt{\frac{\gamma}{D}}(R+\delta)}\right)}{2\gamma\left(e^{-\sqrt{\frac{\gamma}{D}}(R+\delta)} - e^{\sqrt{\frac{\gamma}{D}}(2L-R-\delta)}\right)} e^{-\sqrt{\frac{\gamma}{D}}(R+\delta)},$$

$$a_2 = a_1, \qquad a_6 = a_5 e^{2\sqrt{\frac{\gamma}{D}}L}.$$

Using these constants, one can construct the steady-state solution $u^E(x)$. Note that $u^E \in H_0^1(\Omega)$.

Let $w(x) := u(t,x) - u^E(x)$, then we obtain

$$\begin{cases} \rho w_{tt} - (\mu_1+\mu_2)w_{xxt} - (\lambda+2\mu)w_{xx} + \gamma w = 0, \\ w(t,0) = 0, \quad w(t,L) = 0, \\ w(0,x) = u_0(x) - u^E(x), \quad w_t(0,x) = q_0(x). \end{cases} \tag{21}$$

For the solution of the above equation, we set $w(t,x) = e^{\alpha t}\overline{w}(x)$, to obtain

$$\alpha^2\rho\overline{w}(x) - \alpha(\mu_1+\mu_2)\overline{w}''(x) - (\lambda+2\mu)\overline{w}''(x) + \gamma\overline{w}(x) = 0. \tag{22}$$

Rearrangement gives

$$\overline{w}''(x) - \frac{\gamma+\alpha^2\rho}{(\mu_1+\mu_2)\alpha+\lambda+2\mu}\overline{w}(x) = 0, w(0) = 0, w(L) = 0. \tag{23}$$

Hence $-\frac{\gamma+\alpha_k^2\rho}{(\mu_1+\mu_2)\alpha_k+\lambda+2\mu} = \frac{k^2\pi^2}{L^2} =: \tilde{\lambda}_k^2$ represent eigenvalues with normalized eigenfunctions $\sqrt{\frac{2}{L}}\sin\left(\frac{k\pi x}{L}\right)$, for $k \geq 1$, to the Sturm–Liouville problem. Herewith, solutions are given by

$$\begin{aligned} &w(t,x) = \sum_{k=1}^{\infty} A_k\overline{w}_k^+(t,x) + B_k\overline{w_k^-}(t,x), \text{ where } \overline{w}_k^{\pm}(t,x) = \sqrt{\frac{2}{L}}e^{\alpha_k^{\pm}t}\sin\left(\frac{k\pi x}{L}\right), \\ &\text{where } \alpha_k^{\pm} \text{ follows from } -\frac{\gamma+(\alpha_k^{\pm})^2\rho}{(\mu_1+\mu_2)\alpha_k^{\pm}+\lambda+2\mu} = \tilde{\lambda}_k^2. \end{aligned} \tag{24}$$

The solution for $\alpha_k^{\pm}$ is given by

$$\begin{aligned} \alpha_k^{\pm} &= \frac{-\tilde{\lambda}_k^2(\mu_1+\mu_2) \pm \sqrt{\tilde{\lambda}_k^4(\mu_1+\mu_2)^2 - 4\rho(\gamma+\tilde{\lambda}_k^2(\lambda+2\mu))}}{2\rho}, && \text{if } \rho > 0, \\ \alpha_k &= -\frac{\frac{\gamma}{\tilde{\lambda}_k^2}+\lambda+2\mu}{\mu_1+\mu_2} = \alpha_k^+ = \alpha_k^- < 0, && \text{if } \rho = 0. \end{aligned} \tag{25}$$

If $\rho > 0$, then $\alpha_k^{\pm} < 0$ or $\alpha_k^{\pm} \notin \mathbb{R}$. From the initial conditions, it follows that

$$\begin{aligned} A_k &= \frac{2}{L}\frac{\sqrt{1}}{\alpha_k^- - \alpha_k^+}\int_0^L \left(w(0,\overline{x})\alpha_k^- - w_t(0,x)\right)\sin\left(\frac{k\pi\overline{x}}{L}\right)d\overline{x}, \\ B_k &= \frac{2}{L}\frac{\sqrt{1}}{\alpha_k^+ - \alpha_k^-}\int_0^L \left(w(0,\overline{x})\alpha_k^+ - w_t(0,x)\right)\sin\left(\frac{k\pi\overline{x}}{L}\right)d\overline{x}. \end{aligned} \tag{26}$$

This can be written in a more compact shape in terms of Green's Functions by

$$w(t,x) = \int_0^L G_1(x,\overline{x},0,t)w(0,\overline{x}) - G_2(x,\overline{x},0,t)w_t(0,\overline{x})d\overline{x}, \tag{27}$$

where

$$G_1((x,\overline{x},\overline{t},t) = \frac{2}{L}\sum_{k=1}^{\infty}\left[\frac{\alpha_k^+}{\alpha_k^+-\alpha_k^-}e^{\alpha_k^+(t-\overline{t})} + \frac{\alpha_k^-}{\alpha_k^--\alpha_k^+}e^{\alpha_k^-(t-\overline{t})}\right]\sin\left(\frac{k\pi x}{L}\right)\sin\left(\frac{k\pi\overline{x}}{L}\right),$$

$$G_2((x,\overline{x},\overline{t},t) = \frac{2}{L}\sum_{k=1}^{\infty}\left[\frac{1}{\alpha_k^+-\alpha_k^-}e^{\alpha_k^+(t-\overline{t})} + \frac{1}{\alpha_k^--\alpha_k^+}e^{\alpha_k^-(t-\overline{t})}\right]\sin\left(\frac{k\pi x}{L}\right)\sin\left(\frac{k\pi\overline{x}}{L}\right),$$

provided that $\alpha_k^+ \neq \alpha_k^-$, and $\alpha_k^+,\ \alpha_k^- \in \mathbb{R}$. For $\rho = 0$, similar expressions can be obtained where the initial condition for $\underline{w}_t$ is not used. Note that w is a classical smooth solution. From these considerations, we see that solutions are non-oscillatory if $\alpha_k^+,\ \alpha_k^- \in \mathbb{R}$, that is if

$$\rho \leq \frac{\tilde{\lambda}_k^4(\mu_1+\mu_2)^2}{4(\gamma+\tilde{\lambda}_k^2(\lambda+2\mu))}, \text{ for } k \geq 1. \tag{28}$$

Since the above relation should hold for any $k \geq 1$, we formulate the condition for non-oscillatory convergence to the steady–state equilibrium in the following result:

Theorem 3. *The solution to the problem of equation (18) converges in a non-oscillatory manner to the steady–state equilibrium iff*

$$\rho \leq \frac{\pi^4(\mu_1+\mu_2)^2}{4L^4(\gamma+\frac{\pi^2}{L^2}(\lambda+2\mu))}. \tag{29}$$

Similarly if a Neumann condition is imposed on $x = 0$:

$$\rho \leq \frac{\pi^4(\mu_1+\mu_2)^2}{64L^4(\gamma+\frac{\pi^2}{4L^2}(\lambda+2\mu))}. \tag{30}$$

Considering the following (ranges of) parameter values from [7] (page 429), $\lambda \approx 0.3\,\text{Pa}$, $\mu \approx 0.4\,\text{Pa}$, $\mu_1 \in (25, 7\cdot 10^6)\,\text{Pa/s}$, and $\mu_2 \in (10^4, 10^8)\,\text{Pa/s}$, with reaction spring force constant $\gamma = 0$, it follows that with $\rho \approx 1,050\,\text{kg/m}^3$, the bounds in the above theorem are easily satisfied, and that convergence proceeds monotonically. This implies that the Kelvin–Voigt model, where $\rho = 0$ is used in [3], provides a reasonable approximation. We finally remark that the actual solution $u(t,x)$ is obtained by the addition of a H^1–part and a classical component:

$$u(t,x) = u^E(x) + w(t,x) = u^E(x) + \int_0^L G_1(x,\overline{x},0,t)w(0,\overline{x}) - G_2(x,\overline{x},0,t)w_t(0,\overline{x})d\overline{x}.$$

5 Conclusions

We analyzed a simple visco–elastic model in terms of convergence to the steady–state equilibrium, that is stability was dealt with. Further, we contructed analytic solutions in terms of Green's Functions. Using these analytic solutions, we provide bounds for monotonic convergence for the relation of the physical parameters involved in terms of densities, viscosities and Lamé parameters. Our study also indicates that convergence should be monotonic if realistic biological values for soft tissues are used and that for these system the Kelvin–Voigt model, where inertia is neglected, provides a reasonable approximation. The insights developed in this paper can be used for the sake of validation of numerical solutions and contribute to the understanding of the visco–elastic system.

References

1. Eck, Ch., Jarusek, J.: Existence results for the static contact problem with Coulomb friction. Math. Mod. Meth. Appl. Sci. **8(3)**, 445–468 (1998)
2. Gilbert, R.P., Panchenko, A., Xie, X.: Homogenization of a visco–elastic matrix in linear frictional contact. Math. Meth. Appl. Sci. **28**, 309–328 (2005)
3. Vermolen, F.J., Javierre, E.: A finite-element model for healing of cutaneous wounds combining contraction, angiogenesis and closure. J. Math. Biol. to appear (2012)
4. Evans, L.C.: Partial Differential Equations. Graduate Studies in Mathematics, volume 19, American Mathematical Society, Rhode–Island (1998)
5. Brenner, S.C., Scott, L.R.: The Mathematical Theory of Finite Element Methods. Springer, New York (2008)
6. Braess, D.: Finite Elements: Theory, fast solvers, and applications in solid mechanics. Cambridge University Press, Cambridge (2001)
7. Murray, J.D.: Mathematical Biology, II Spatial Models and Biomedical Applications. Springer, New York (2003)
8. Adam, J.A.: A simplified model of wound healing (with particular reference to the critical size defect). Math. and Comput. Modell., **30**, 23–22 (1999)

Extending the Volume of Fluid Method to Higher Order Accuracy

J.C.G. Verschaeve

Abstract In the present discussion we present a combination of the volume of fluid method with the front tracking method for the advection of the interface between two immiscible phases in two dimensions. The mass of each phase is conserved up to roundoff accuracy. The interface is thereby represented very accurately. A drawback of the present method is that topological changes are not handled 'automatically' as for the volume of fluid method.

1 Introduction

Two-phase flow can be found in many industrial applications. Two popular methods for the computation of the evolution of the interface are the volume of fluid method [3] and the front tracking method [6]. The volume of fluid method is an interface capturing method, meaning that the interface is not tracked explicitly but obtained by means of another quantity, the volume fraction. The front tracking method, on the other hand, follows the interface in a Lagrangian way by means of marker particles, distributed along the interface and advected by the flow. An advantage of the volume of fluid method is that the volume of each phase is conserved during simulation. However, geometric quantities such as position, normals and curvature are relatively difficult to obtain by means of the volume fractions which leads to less accurate results than for the front tracking method. In order to improve mass conservation for the front tracking method Ye et al. [7] propose a correction scheme, correcting the position of the marker particles, such that less volume is lost.

In the present discussion, we combine the volume of fluid method and the front tracking method in two dimensions in order to generate a method which accurately tracks the interface while conserving the volume of each phase. A drawback of the

J.C.G. Verschaeve (✉)
NGI, Postboks 3930 Ullevaal Stadion, 0806, Oslo, Norway
e-mail: joris.verschaeve@ngi.no

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_65,

method is that topological changes such as coalescence and break up are not handled 'automatically'. The interface in the present discussion is divided, similar to the segment projection method [5], into segments which can be described by a function taking one of the coordinates as argument and returning the interface position of the other. However, in order to guarantee volume conservation, the segment, in the present discussion, is not described by a function returning the position of the interface but by the primitive of this function, which will be called the segment primitive function. The advection of this segment primitive function is then done by volume tracking.

The remainder of this discussion is organized as follows. In Sect. 2 the derivation of the present method is presented. The numerical verification is then done in Sect. 3, whereas the conclusions are drawn in Sect. 4.

2 Derivation

For the computation of the area of a regular domain $\Omega \subset \mathbb{R}^2$ the following method will often be used. Since Ω is regular, its boundary $\partial\Omega$ can be represented by a set of parametrizations γ_r, $r = 1, \ldots, R$:

$$\begin{aligned} \gamma_r : [a_r, b_r] &\to \Omega \\ s &\mapsto \mathbf{x}_r(s) = \begin{pmatrix} x_r(s) \\ y_r(s) \end{pmatrix}. \end{aligned} \tag{1}$$

A normal vector $\mathbf{n}_r^*$ on $\partial\Omega$ can be defined by:

$$\mathbf{n}_r^* = \pm \begin{pmatrix} y_r'(s) \\ -x_r'(s) \end{pmatrix}, \tag{2}$$

where the sign $\pm$ is chosen such that $\mathbf{n}_r^*$ points into the exterior of Ω. The area $\mathrm{vol}(\Omega)$ of Ω can then be computed by, cf. Arvo [1]:

$$\mathrm{vol}(\Omega) = \int_\Omega dV \tag{3}$$

$$= \int_\Omega \frac{1}{2} (\nabla \cdot \mathbf{x})\, dV \tag{4}$$

$$= \frac{1}{2} \int_{\partial\Omega} \mathbf{x} \cdot \mathbf{n}\, dS \tag{5}$$

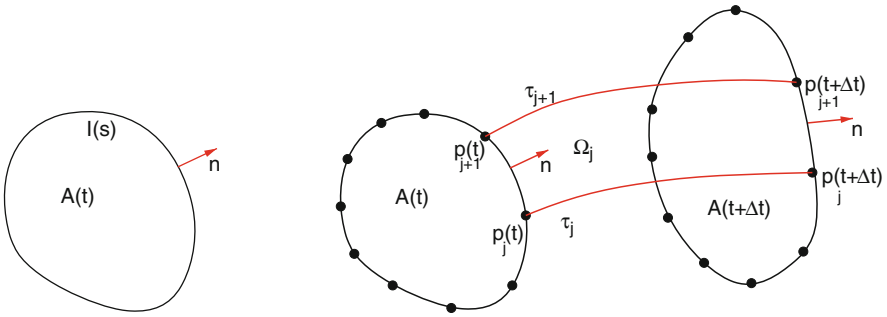


Fig. 1 Advection of the drop. *Left*: the interface of the drop $A(t)$ is described by a parametrization $I(s)$. *Right*: marker particles distributed along the interface are traced along their trajectories τ to their positions at $t + \Delta t$

$$= \sum_{r=1}^{R} \pm \underbrace{\frac{1}{2} \int_{a_r}^{b_r} x_r(s) y_r'(s) - y_r(s) x_r'(s)\, ds}_{=S_r}, \tag{6}$$

$$= \sum_{r=1}^{R} \pm S_r \tag{7}$$

where the Gauss theorem has been used and the definite line integral is denoted by the uppercase letter S. In the following the letter S with subscript shall be used to denote definite integrals of this form along a line. Indefinite integrals of this form shall be marked by their functional dependence, i.e. $S(s)$, meaning:

$$S(s) = \frac{1}{2} \int_{0}^{s} x(\sigma) y'(\sigma) - y(\sigma) x'(\sigma)\, d\sigma, \tag{8}$$

where $(x(s), y(s))$ are the coordinates of a parametrization.

In the present case, the interface of a drop at time t is described by a parametrization $I(s) : s \in [0, 2\pi) \mapsto (x(s), y(s))$, cf. Fig. 1. Similar to the front tracking method, marker particles p_j, $j = 1, N$ with position $\mathbf{x}_j = (x_j, y_j) = (x(s_j), y(s_j))$, are distributed along the interface. Tracing the marker particles forward along the trajectories τ_j by solving,

$$\dot{\mathbf{x}_j} = \mathbf{u}(\mathbf{x}_j, t), \tag{9}$$

allows us to find the position of the marker particles at time $t + \Delta t$ [6], cf. Fig. 1.

In order to obtain a mass conserving method for the advection of the interface, the set Ω_j, sketched in Figs. 1 and 2, shall be of use in the following. This set

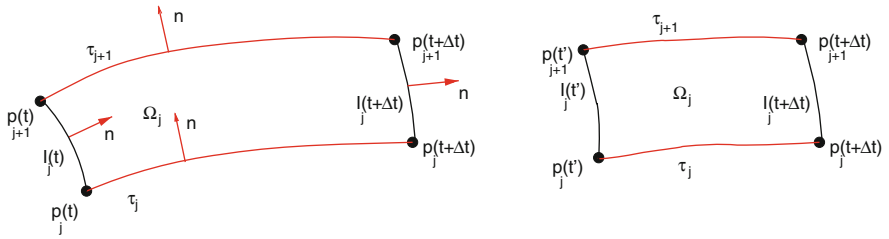


Fig. 2 The set Ω_j. *Left*: at time t. *Right*: at time t' ($t \leq t' \leq t + \Delta t$)

Ω_j is bounded by two trajectories τ_j and τ_{j+1} of two consecutive marker particles and by a section $I_j(t)$ from $p_j(t)$ and $p_{j+1}(t)$ of the interface at time t and the corresponding section $I_j(t + \Delta t)$ of the interface at time $t + \Delta t$, cf. Fig. 2. By Eq. (7), the area of Ω_j can be computed as:

$$\mathrm{vol}(\Omega_j) = S(s_j, t) - S(s_{j+1}, t) + S_{\tau_j} - S_{\tau_{j+1}} + S(s_{j+1}, t + \Delta t) - S(s_j, t + \Delta t). \tag{10}$$

The quantity $S(s,t)$ is the indefinite integral over the interface at time t and defined by Eq. (8). The quantity S_{τ_j} is the indefinite integral over the trajectory τ_j and defined by Eq. (6). The last term, $S(s, t + \Delta t)$, is the indefinite integral over the interface at time $t + \Delta t$. The normals of the parametrizations, cf. Fig. 2, determine the signs in front of the terms in Eq. (10). We remark that $\mathrm{vol}(\Omega_j)$ as defined by Eq. (10) can be negative or positive. This is not a flaw of the method, but is necessary in order to have mass conservation. Equation (10) can be solved for the values of $S(s_j, t + \Delta t)$ at the new time $t + \Delta t$:

$$S(s_{j+1}, t + \Delta t) - S(s_j, t + \Delta t) = S(s_{j+1}, t) - S(s_j, t) - S_{\tau_j} + S_{\tau_{j+1}} + \mathrm{vol}(\Omega_j). \tag{11}$$

Equation (11) allows us to obtain the values of $S(s, t + \Delta t)$ at s_{j+1} and s_j, respecively, at the new time $t + \Delta t$ as a function of the old values $S(s,t)$ at s_{j+1} and s_j, respectively, of the definite integrals S_{τ_j} and $S_{\tau_{j+1}}$, over the trajectories τ_j and τ_{j+1}, respectively, and of the area $\mathrm{vol}(\Omega_j)$. The definite integrals S_{τ_j} and $S_{\tau_{j+1}}$,respectively, can be computed by a quadrature formula on Gauss-Lobatto-Legendre nodes of the interval $[t, t + \Delta t]$. The quantity $\mathrm{vol}(\Omega_j)$, on the other hand, is related to the velocity field $\mathbf{u}(\mathbf{x}, t)$ advecting the drop. For a point $t' \in [t, t + \Delta t]$ in time the set $\Omega_j(t')$, bounded by the trajectories τ_j and τ_{j+1}, the section of the interface $I_j(t + \Delta t)$ at time $t + \Delta t$ and the section of the interface $I_j(t')$ now at time t' has decreased, cf. Fig. 2. When taking $t' = t + \Delta t$ the area of $\Omega_j(t')$ is actually zero, since the interface has then moved to its position at $t + \Delta t$. For an incompressible flow, the change of $\mathrm{vol}(\Omega_j)(t')$ is given by:

$$\frac{d}{dt'} \int_{\Omega_j} d\Omega = - \int_{\partial\Omega_j} \mathbf{u} \cdot \mathbf{n} \, d\sigma. \tag{12}$$

The right hand side corresponds to the volume flow leaving and entering $\Omega_j(t')$ and is explicitly given by:

$$\frac{d}{dt'}\int\limits_{\Omega_j} d\Omega = -\int\limits_{I(t')} \mathbf{u}_r \cdot \mathbf{n}\, d\sigma - \int\limits_{\tau_j} \mathbf{u}\cdot\mathbf{n}\, d\sigma - \int\limits_{\tau_{j+1}} \mathbf{u}\cdot\mathbf{n}\, d\sigma - \int\limits_{I(t+\Delta t)} \mathbf{u}\cdot\mathbf{n}\, d\sigma$$

$$\approx -\underbrace{\int\limits_{I(t+\Delta t)} \mathbf{u}\cdot\mathbf{n}\, d\sigma}_{=Q_j}. \tag{13}$$

The volume flow over the moving interface vanishes, since the relative velocity $\mathbf{u}_r$ is zero. In addition, the flow over the trajectories is neglected, since the trajectory will be close to a streamline for small Δt. Integrating equation (13) in time from t to $t + \Delta t$ leads to

$$\mathrm{vol}(\Omega_j) = -\int\limits_{\Omega_j(t+\Delta t)} d\Omega + \int\limits_{\Omega_j(t)} d\Omega = \int\limits_{t}^{t+\Delta t} Q_j(t')\, dt', \tag{14}$$

which gives us the result that the area $\mathrm{vol}(\Omega_j)$ is approximately given by the volume which, in the time interval $[t, t + \Delta t]$, flowed over the position of the interface at time $t + \Delta t$. We remark that neglecting the flow over the trajectories does not pose a problem for mass conservation, since they cancel for two consecutive sets Ω_j. It is, however, necessary, that the sum of Q_j is zero. This is in general not the case, since the discretization of $\mathbf{u}$ produces velocity fields which are not divergence free. If $\epsilon = \sum_{k=0}^{N-1} Q_k$, then the sum of $Q_j^* = Q_j - \epsilon/(N-1)$ vanishes. Therefore Q_j is replaced by Q_j^* in Eq. (14). Thus, we obtain for $S(s_{j+1}, t + \Delta t)$:

$$S(s_{j+1}, t+\Delta t) - S(s_j, t+\Delta t) = S(s_{j+1}, t) - S(s_j, t) - S_{\tau_j} + S_{\tau_{j+1}} + \int\limits_{t}^{t+\Delta t} Q_j^*(t')\, dt'. \tag{15}$$

Having found the quantity $S(s_j, t + \Delta t)$ in function of different quantities, how do we compute the position of the interface, knowing $S(s_j, t + \Delta t)$? In order to solve this problem the interface is divided into segments, which can be described by a function f_r, $r = 0, \ldots, R$, of one of the coordinates and returning the other, cf. Fig. 3. This is the foundation of the segment projection method derived by Tornberg and Engquist [5]. For the present method the segment is not described by f_r but by F_r which is the primitive of f_r, i.e. $f_r = F_r'$. For a segment aligned along the x axis, the position on the interface is therefore given by:

$$\big(x, y\big) = \big(x, f(x)\big) = \big(x, F'(x)\big). \tag{16}$$

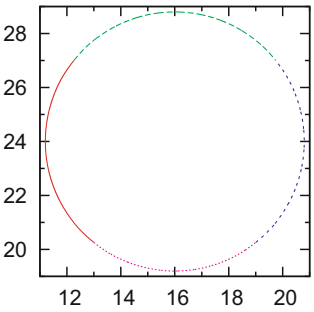


Fig. 3 A circular interface is divided into four segments. Each segments can be described by a function taking either the x or y coordinate as an argument. The segments end where the normal on the interface is inclined by 45° with respect to the grid axes

On the other hand F can be related to S, Eq. (8). Assuming that the segment along x extends from the point $(x(s_{i_r}), y(s_{i_r}))$ to the point $(x(s_{i_r+N_r}), y(s_{i_r+N_r}))$, we have

$$S(s_{i_r+1}) - S(s_{i_r}) = \int_{s_{i_r}}^{s_{i_r+1}} x(s)y'(s) - x'(s)y(s)\,ds \tag{17}$$

$$= \int_{x_{i_r}}^{x_{i_r+1}} xF''(x) - F'(x)\,dx \tag{18}$$

$$= x_{i_r+1}y_{i_r+1} - x_{i_r}y_{i_r} - 2\left(F(x_{i_r+1}) - F(x_{i_r})\right) \tag{19}$$

Together with Eq. (15) a set of interpolation points $(x_{i_r+j}, F(x_{i_r+j}))$, $j = 0, \ldots, N_r$ can be found which are then interpolated by means of cubic splines [2]. The points on the interface, where the segments concatenate are those points with a normal inclined by approximately 45° with the grid axes, cf. Fig. 3 .

3 Numerical Verification

As benchmark test, the deformation field test [4] was chosen. The initial set up of this test consists of a square computational domain of side length 1 with a circular drop whose center is located at $x_0 = 0.5, y_0 = 0.5$ and its radius is $r_0 = 0.25$. A velocity field given by means of the following stream function causes the drop to deform, cf. Fig. 4.

$$\psi(x, y, t) = \frac{1}{4\pi}\cos\left(\frac{\pi t}{T}\right)\sin\left(4\pi\left(x + \frac{1}{2}\right)\right)\cos\left(4\pi\left(y + \frac{1}{2}\right)\right), \tag{20}$$

The drop reaches its maximum deformation at time $t = T/2$, when the flow is reversed and the drop returns to its initial position at time $t = T$. The discrepancy

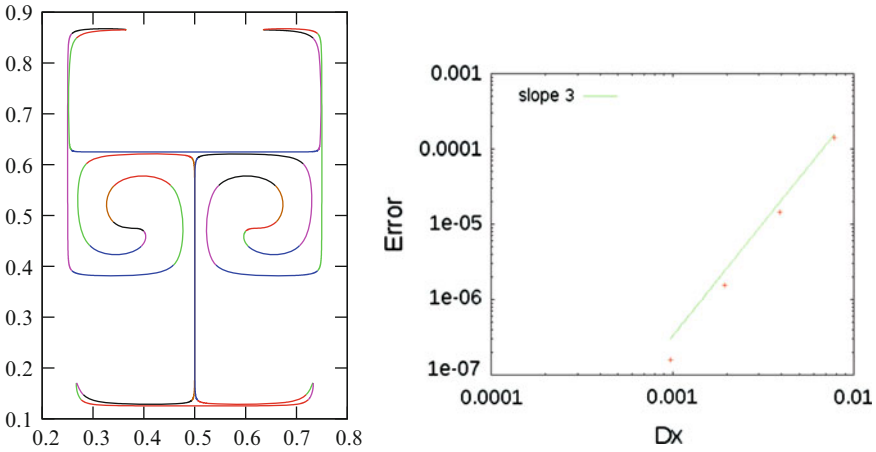


Fig. 4 *Left*: position of the interface for the deformation field test at $t = T/2$. *Right*: error decrease in function of the average distance between marker points

between the exact circle and the numerical result at time $t = T$ serves as a measure of the numerical error. Figure 4 shows the position of the interface for $t = T/2$. We remark that the marker points on the interface are sampled at each time step. This allows to adapt the points depending on the size of the segment. The error, cf. Fig. 4, displays third order convergence with respect to the average distance of the marker points, corresponding to the theoretical order of convergence for the cubic spline [2] since we are measuring the error on the first derivative of the segment primitive. The area of the drop is conserved up to round off accuracy during simulation.

4 Conclusions

In the present discussion, we derived a method to accurately compute the evolution of the interface between two immiscible fluids while conserving the mass of each phase exactly. The interface was represented by segment primitive functions, which is closely related to the segment primitive method [5]. The advection of the interface is done by volume tracking. The resulting method has been tested for the deformation field test [4] and the resulting order of convergence corresponded to the theoretically expected one. Several important issues need still to be investigated. In order to observe the behavior of the method when faced with a non ideal velocity field, the next step is to couple the method to a Navier-Stokes solver. Since the method is related to the volume of fluid method, a scheme might be derived allowing topological changes to occur in a volume of fluid framework. The three dimensional extension of the method is far from being straightforward.

References

1. J. Arvo, editor. *Graphics Gems II*. Morgan Kaufmann Academic Press, San Diego, 1991.
2. G. Micula and S. Micula. *Handbook of Splines*. Kluwer Academic Publishers, 1999.
3. R. Scardovelli and S. Zaleski. Direct numerical simulation of free surface and interface flow. *Annual Review of Fluid Mechanics*, 31:567, 1999.
4. P. Smolarkiewicz. The multi-dimensional crowley advection scheme. *Month. Weather Rev.*, 110:1968–1983, 1982.
5. A.-K. Tornberg and B. Engquist. The segment projection method for interface tracking. *Communications on Pure and Applied Mathematics*, LVI:47–79, 2003.
6. G. Tryggvason, B. Brunner, A. Esmaeeli, D. Juric, N. Al-Rawahi, W. Tauber, J. Han, S. Nas, and Y.-J. Jan. A front-tracking method for the computations of multiphase flow. *Journal of Computational Physics*, 169:708–759, 2001.
7. T. Ye, W. Shyy, and J. N. Chung. A fixed-grid, sharp-interface method for bubble dynamics and phase change. *Journal of Computational Physics*, 174:781–815, 2001.

Stability Estimates and Numerical Comparison of Second Order Time-Stepping Schemes for Fluid-Structure Interactions

T. Wick

Abstract It is well-known that the Crank-Nicolson scheme for pure fluid problems suffers from stability for computations over long-term time intervals. In the presence of fluid-structure interaction in which the fluid equations are reformulated with the help of arbitrary Lagrangian-Eulerian (ALE) mapping, the ALE convection also causes stability problems. In this study, we derive a stability estimate of a monolithically coupled time-discretized fluid-structure interaction problem. Moreover, a numerical comparison of all relevant second order time-stepping schemes, such as secant and tangent Crank-Nicolson, shifted Crank-Nicolson, and Fractional-Step-Theta, is demonstrated. The numerical experiments are based on a benchmark configuration for fluid-structure interactions.

1 Introduction

It is already well-known from pure fluid problems on fixed meshes, that the second order ordinary Crank-Nicolson scheme suffers from instabilities, particularly for long-term computations [7]. Optimal error estimates Crank-Nicolson scheme.

The normally unconditionally stable Crank-Nicolson scheme is restricted by the condition

$$k \leq ch^{2/3}, \tag{1}$$

where k and h denote the time-step size and the mesh-size parameter, respectively. However, the scheme can be stabilized by moving the θ-parameter (using a One-Step-θ scheme, see, e.g., [14]) slightly to the implicit side, leading to the shifted Crank-Nicolson scheme [10, 12]. On the other hand, several authors detected

T. Wick (✉)
Institute of Applied Mathematics, University of Heidelberg, Heidelberg, Germany
e-mail: thomas.wick@iwr.uni-heidelberg.de

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_66,

numerical instabilities on moving domains for higher order time-stepping schemes caused by the ALE convection term [2,4,5,9].

Specifically, the ALE convection term is a numerical artifact that only appears on moving domains [4,5,11]. However, the relevance in numerical computations is not yet completely understand. The stability is closely related to the verification of the *Geometric Conservation Law* (GCL) [2,4,5,9]. Moreover, they proved that the GCL condition does not degrade the accuracy of the numerical schemes. In this study, the previously mentioned results are combined with estimates for structural interactions. Finally, some second-order time-stepping schemes are compared within a numerical study.

2 The Equations

For the fluid, we consider a time-dependent domain $\Omega_f(t) \subset R^d$, $d = 2$, with the boundary $\Gamma = \Gamma_{\text{in}} \cup \Gamma_{\text{wall}} \cup \Gamma_{\text{out}} \cup \Gamma_i$. The boundary part Γ_i denotes later the interface between the fluid subsystem and the structural system. Moreover, we denote by $I = (0, T]$ the time interval. The unknowns are the fluid velocity $v_f : \Omega_f \times R^+ \to R^d$, and the fluid pressure $p_f : \Omega_f \times R^+ \to R$. Then, the Navier-Stokes equations of an incompressible, isothermal fluid read:

$$\begin{aligned} \rho_f \hat{\partial}_t v_f + \rho_f (v_f - w) \cdot \nabla v_f - \text{div}\sigma_f = 0 \quad &\text{in } \Omega_f(t) \times I, \\ \text{div} v_f = 0 \quad &\text{in } \Omega_f(t) \times I, \end{aligned} \tag{2}$$

where w denotes the fluid domain velocity, which is defined by $v_f = w = \partial \hat{u}_s$ on Γ_i. The fluid Cauchy stress tensor reads

$$\sigma_f := -p_f I + \mu_f (\nabla v_f + \nabla v_f^T).$$

The (dynamic) viscosity is denoted by $\mu_f := \rho_f \nu_f$ in which ρ_f and ν_f denote density and the (kinematic) viscosity, respectively. We notice that the term $\hat{\partial}_t v_f$ denotes the ALE time derivative [6].

The structure problem is defined in a fixed domain $\widehat{\Omega}_s$ with the boundary $\widehat{\Gamma}_{\text{s,fixed}} \cup \widehat{\Gamma}_i$. The structure is fixed on $\widehat{\Gamma}_{\text{s,fixed}}$ using homogenous Dirichlet conditions. The physical unknown is the structure displacement $\hat{u}_s : \widehat{\Omega}_s \times R^+ \to R^3$. The governing equations for the structural subsystem in a mixed formulation read ([15]):

$$\begin{aligned} \hat{\rho}_s \partial_t \hat{v}_s - \widehat{\text{div}}(\widehat{F}\,\widehat{\Sigma}_s(\hat{u}_s)) + \gamma_w \hat{v}_s - \gamma_s \widehat{\text{div}}(\hat{\epsilon}(\hat{v}_s)) = \hat{\rho}_s \hat{f}_s \quad &\text{in } \widehat{\Omega}_s,\ t \in I, \\ \hat{\rho}_s (\partial_t \hat{u}_s - \hat{v}_s) = 0 \quad &\text{in } \widehat{\Omega}_s,\ t \in I, \\ \hat{u}_s = 0, \quad \hat{v}_s = \partial_t \hat{u}_s = 0 \quad &\text{on } \widehat{\Gamma}_{s,\text{in,out,wall}} = \widehat{\Gamma}_{\text{fixed}}. \end{aligned} \tag{3}$$

The structural stress tensor (namely the Saint Venant Kirchhoff material - STVK), $\widehat{\Sigma}_s$, is defined as

$$\widehat{\Sigma}_s = (\lambda_s(\mathrm{tr}\widehat{E})\hat{I} + 2\mu_s\widehat{E}), \quad \widehat{E} = 2^{-1}(\widehat{F}^T\widehat{F} - \hat{I}),$$
$$\widehat{F} = \hat{I} + \widehat{\nabla}\hat{u}_s, \quad \hat{\epsilon}(\hat{v}_s) = 2^{-1}(\widehat{\nabla}\hat{v}_s + \widehat{\nabla}\hat{v}_s^T),$$

in which $\hat{I}$ denotes the identity tensor. The elastic structure is characterized by the Lamé coefficients μ_s, λ_s.

3 Stability of the Time-Discretized ALE Fluid Problem and the FSI Problem

In this section, a slight modification of the classical Crank-Nicolson scheme (i.e., it is a Gauss-Legendre implicit second-order Runge-Kutta method) is considered [5]. We work with an ALE map $\hat{\mathscr{A}}$ which is defined from the previous time step t_{n-1} to the the present time step t_n. Thus, the reference configuration at time step t_n is denoted by Ω^n. Moreover, $v^n \in \Omega^n$ is used as an approximation to $v(t_n)$, which is transported from Ω^n to any other configuration Ω^l (for $l \neq n$) through the ALE map ([11]):

$$\hat{\mathscr{A}}_{n,l} = \hat{\mathscr{A}}_l \circ \hat{\mathscr{A}}_n^{-1}.$$

For the sake of notation, we omit the explicit representation of the ALE map when we work with the value v^n in a domain Ω_l with $n \neq l$, i.e.,

$$\int_{\Omega_l} v^n \,\mathrm{d}x := \int_{\Omega_l} v^n \circ \hat{\mathscr{A}}_{n,l} \,\mathrm{d}x, \quad \text{and} \quad ||v^n||_{\Omega^l} := ||v^n \circ \hat{\mathscr{A}}_{n,l}||_{\Omega^l},$$

which we use frequently in the following.

To get a stability result for the time-discretized Crank-Nicolson scheme on moving domains, we use the methodology used in [4, 5, 11]. It holds:

Proposition 1. *For the time-discretized solution of ALE fluid problems with the help of the Crank-Nicolson scheme holds:*

$$\rho_f||v_f^{n+1}||^2_{\Omega_f^{n+1}} + k\mu_f||D(v_f^{n+1} + v_f^n)||^2_{\Omega_f^{n+1}} + \frac{k}{4}\rho_f \int_{\Omega_f^{n+1}} \nabla \cdot w^{n+\frac{1}{2}} \, |v_f^{n+1} + v_f^n|^2 \,\mathrm{d}x$$
$$= \rho_f||v_f^n||^2_{\Omega_f^{n+1}}.$$

For $\nabla \cdot w > 0$ for all $x \in \Omega_f$ and for all $t \in I$ (a uniform contraction of the mesh), the Crank-Nicolson scheme is unconditionally stable. Otherwise, the ALE convection term causes instabilities that restricts the choice of the time step size [5, 11]. Therefore, the convection term is estimated as follows:

$$\frac{k\rho_f}{4}\int_{\Omega_f^{n+1}} \nabla\cdot w^{n+1}\,|v_f^{n+1}+v_f^n|^2\,\mathrm{d}x \le k\delta_w\left(||v_f^{n+1}||^2_{\Omega_f^{n+1}}+||v_f^n||^2_{\Omega_f^{n+1}}\right),$$

in which the Young inequality is used to estimate the right-hand-side term. Specifically, it holds:

$$k \le \delta_w^{-1}, \quad \textit{with} \quad \delta_w := \frac{\rho_f}{2}||\nabla\cdot w^{n+1}||_{L^\infty(\Omega_f^{n+1})}.$$

Proof. For the proof, we refer the reader to [15].

Combining this result with the restriction (1), which was analyzed by Rannacher et al. [7, 12], provides us

Proposition 2. *Using the ordinary (i.e., unstabilized) Crank-Nicolson scheme leads to the following time step condition for pure fluid problems on moving domains:*

$$k \le \min\left\{ch^{2/3}, \delta_w^{-1}\right\}. \tag{4}$$

Using the shifted Crank-Nicolson scheme [12], the first condition in (4) can be removed, such that $k \le k^*$, with some constant k^* that only depends on the problem.

It seems that the time step restriction $k \le \delta_w^{-1}$ induced by the mesh movement seems to be of lower order and it has less influence than the first condition $k \le ch^{2/3}$. In fact, Formaggia and Nobile [5], p. 4098, state that they found no example of blow-up caused by the ALE convection term for linear advection-diffusion equations. This might be due to the fact that the ALE convection term is only defined on a lower-dimensional manifold and not over the whole domain.

We utilize the previous results to analyze the monolithically coupled fluid-structure interaction system. First, we recall the coupling conditions that are required for an implicit solution algorithm:

$$\hat{u}_f^{n+1} = \hat{u}_s^{n+1} \quad \text{on } \widehat{\Gamma}_i, \qquad \hat{w}^{n+1} = \frac{1}{k}(\hat{u}_f^{n+1}-\hat{u}_f^n) \quad \text{in } \widehat{\Omega}_f, \tag{5}$$
$$2^{-1}(v_f^{n+1}+v_f^n) = w_f^{n+1} \quad \text{on } \Gamma_i, \qquad \hat{u}_s^{n+1} = 0 \quad \text{on } \widehat{\Gamma}_{s,D}.$$

Using the Crank-Nicolson scheme for temporal discretization, the second relation in (5), can be further developed into

$$\hat{w}^{n+1} = k^{-1}(\hat{u}_f^{n+1}-\hat{u}_f^n) = 2^{-1}(\hat{v}_f^{n+1}+\hat{v}_f^n). \tag{6}$$

Fernández and Gerbeau [3] proved a result using the backward Euler scheme to discretize the fluid. The structure is discretized with a second-order mid-point rule. In our study, both systems are time-discretized with the same time-stepping scheme. We emphasize, that fluid flows on moving meshes with a Crank-Nicolson

time discretization only serve for a conditioned stability (see Proposition 1). Consequently, we cannot expect a better result for the overall problem.

To derive the next proposition, we use a Crank-Nicolson discretization for the fluid with the stability result proven in Proposition 1. The coupling term on the interface Γ_i reads:

$$\sigma_f(v_f^{n+1} + v_f^n)n_f + \widehat{F}\,\widehat{\Sigma}(\hat{u}_s^{n+1} + \hat{u}_s^n)\hat{n}_s + \gamma_s\hat{\epsilon}(\hat{v}_s^{n+1} + \hat{v}_s^n)\hat{n}_s = 0. \tag{7}$$

Proposition 3. *Let the fluid-structure interaction problem be coupled via an implicit solution algorithm and let both subproblems be time-discretized with the second order Crank-Nicolson scheme. The coupled problem is assumed to be isolated, i.e.,* $v_f^{n+1} = 0$ *on* $\partial\Omega_f \setminus \Gamma_i$ *and* $\widehat{F}\,\widehat{\Sigma}(\hat{u}_s^{n+1})\hat{n}_s = 0$ *on* $\partial\widehat{\Omega}_s \setminus \widehat{\Gamma}_i$. *Further, in the case of strong damping* $\gamma_w > 0$, *let* $\hat{\epsilon}(\hat{v}_s^{n+1})\hat{n}_s = 0$ *on* $\partial\widehat{\Omega}_s \setminus \widehat{\Gamma}_i$. *Then,*

$$\begin{aligned}
&\rho_f||v_f^{n+1}||^2_{\Omega_f^{n+1}} + \hat{\rho}_s||\hat{v}_s^{n+1}||^2_{\widehat{\Omega}_s} + \int_{\widehat{\Omega}_s} W\big(\widehat{F}(\hat{u}_s^{n+1})\big)\,\mathrm{d}x + k\mu_f||D(v_f^{n+1} + v_f^n)||^2_{\Omega_f^{n+1}} \\
&\quad + \frac{k\rho_f}{4}\int_{\Omega_f^{n+1}} \nabla\cdot w^{n+1}|v_f^{n+1} + v_f^n|^2\,\mathrm{d}x + \frac{k\gamma_w}{2}||\hat{v}_s^{n+1}||^2_{\widehat{\Omega}_s} + \frac{k\gamma_s}{2}||\hat{\epsilon}(\hat{v}_s^{n+1})||^2_E \\
&\leq \rho_f||v_f^n||^2_{\Omega_f^{n+1}} + \rho_s||\hat{v}_s^n||^2_{\widehat{\Omega}_s} + \int_{\widehat{\Omega}_s} W(\widehat{F}(\hat{u}_s^n))\,\mathrm{d}x + \frac{k\gamma_w}{2}||\hat{v}_s^n||^2_{\widehat{\Omega}_s} + \frac{k\gamma_s}{2}||\hat{\epsilon}(\hat{v}_s^n)||^2_E.
\end{aligned}$$

Proof. For the proof, we refer to [15].

Comparing Propositions 1 and 3, we notice that global stability of solutions depends only on the uncertainty of the ALE convection term. We draw the following conclusion from our previous findings:

Hypothesis 1 (Stable long-term computations of FSI problems). *Numerically stable long-term computations of fluid-structure interaction can be computed by (at least) strictly A-stable time-stepping schemes (such as the shifted Crank-Nicolson scheme and the Fractional-Step-θ scheme) provided that the time step k is restricted by*

$$k \leq \delta_w^{-1},$$

as shown in Proposition 2.

3.1 Discretization of the ALE Convection Term

In this section, we discuss possible temporal discretizations of the ALE convection term. From Eq. (2), we extract

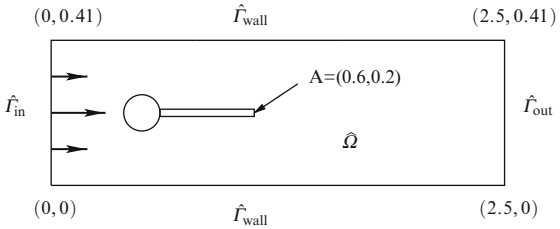


Fig. 1 Elastic beam attached at a cylinder with *circle-center* $C = (0.2, 0.2)$ and radius $r = 0.05$

$$(v_f - w) \cdot \nabla v_f = v_f \cdot \nabla v_f - w \cdot \nabla v_f \quad \text{in } \Omega_f.$$

In detail, the One-Step-θ discretization yields ($\theta = 0.5$ or $\theta = 0.5 + k$):

$$\theta v_f \cdot \nabla v_f + (1-\theta) v_f^{n-1} \cdot \nabla v_f^{n-1} - k^{-1}(u_f - u_f^{n-1}) \cdot \nabla v_f, \quad \text{[Sec } CN(v)\text{]}$$

$$\theta v_f \cdot \nabla v_f + (1-\theta) v_f^{n-1} \cdot \nabla v_f^{n-1} - \theta w \cdot \nabla v_f - (1-\theta) w^{n-1} \cdot \nabla v_f^{n-1}, \quad \text{[Sec } CN(vw)\text{]}$$

$$\theta(v_f + v_f^{n-1}) \cdot \nabla \theta(v_f + v_f^{n-1}) - \theta(w + w^{n-1}) \cdot \nabla \theta(v_f + v_f^{n-1}) \quad \text{[Tang } CN(vw)\text{]}.$$

We notice that the tangential scheme is used for a stability and accuracy analysis for pure fluid problems [7]. This scheme is slightly more stable than the secant Crank-Nicolson scheme [13], which we also observed in our numerical tests (see at left of Fig. 2).

4 Numerical Tests and Observations

In the final section, we compare all relevant second-order time-stepping schemes for solving fluid-structure interaction. For details on temporal discretization, we refer the reader to [15, 16]. Spatial discretization is based on a Galerkin finite element scheme; for details on our solution algorithm, we refer to [15, 16].

We consider the numerical benchmark test FSI 2 [1, 8]. The (qualitative) convergence with respect to space and time on three different (globally-refined) mesh levels is studied using with 1914, 7176 and 27744 degrees of freedom using the Q_2^c/P_1^{dc} element. Moreover, we use three different time levels with the time steps $k = 0.01, 0.005$ and 0.001. It is sufficient to study the results for the drag evaluation because we observed the same qualitative behavior for all the four quantities of interest (the x- and the y-displacement, the drag, and the lift). Specifically, the drag is computed as line integral over the cylinder and the interface of the elastic beam. The configuration is sketched in Fig. 1.

Boundary conditions

A parabolic inflow velocity profile is given on $\hat{\Gamma}_{\text{in}}$ by

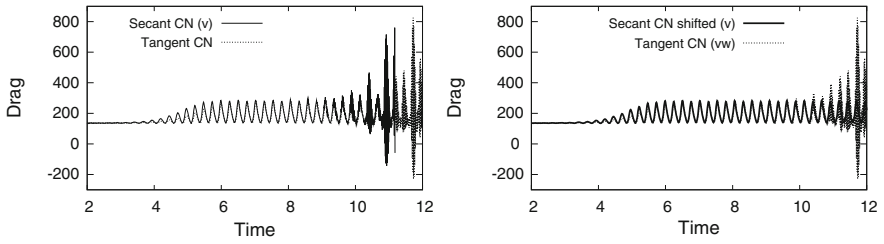


Fig. 2 Blow-up (using the time step $k = 0.01$) of the unstabilized Crank-Nicolson schemes (secant and tangent) whereas the shifted Crank-Nicolson schemes is stable throughout the whole time interval. We notice that the secant Crank-Nicolson scheme exhibits the instabilities earlier than the tangent version. The unit of the time axis is s, whereas the drag unit is $\mathrm{kg/m\,s^2}$

$$v_f(0, y) = 1.5\bar{v}\frac{4y(H - y)}{H^2}, \quad \bar{v} = 1.0ms^{-1}.$$

At the outlet $\hat{\Gamma}_{\text{out}}$ the do-nothing outflow condition is imposed. Homogenous Dirichlet boundary conditions are prescribed on the remaining boundary parts.

Initial conditions

For the unsteady tests, a smooth increase of the velocity-profile in time is chosen:

$$v_f(t; 0, y) = \begin{cases} v_f(0, y)\frac{1-\cos(\frac{\pi}{2}t)}{2} & \text{if } t < 2.0s \\ v_f(0, y) & \text{otherwise.} \end{cases}$$

Parameters

We choose for our computation the following parameters. For the fluid, we use $\mu_f = \mathrm{m^2 s^{-1}}$. The elastic structure is characterized by $\varrho_s = 10^4 \mathrm{kgm^{-3}}$, $\nu_s = 0.4$, $\mu_s = 5 * 10^5 \mathrm{kgm^{-1}s^{-2}}$. Moreover, we set $\gamma_w = \gamma_s = 0$.

Discussion of the results

We observed in our computations that there are only minor differences in the drag evaluation computed with the unstabilized Crank-Nicolson scheme using the different ALE convection term discretizations defined in the problems above. Specifically, we observed unstable behavior (blow-up) for computations over long-term intervals, as illustrated in Fig. 2. Naturally, we expected this behavior from our previous numerical analysis.

As expected, the shifted Crank-Nicolson scheme and the Fractional-Step-θ scheme showed no stability problems in long-term computations, even for the large time step $k = 0.01$ (see at left of Fig. 3). This result indicates that the instabilities induced by the ALE convection term have minor consequences, and our observation is in agreement with the statement in [5]. Furthermore, all time-stepping schemes are stable over the entire time interval for a sufficiently small time step $k = 0.001$; (see the bottom Fig. 3). Consequently, we were able to find a suitable bound such that the requirements of Proposition 2 are satisfied.

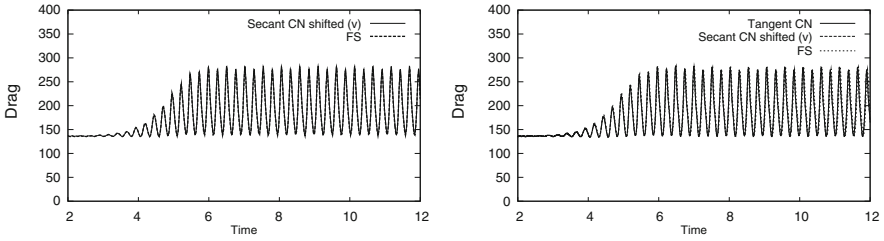


Fig. 3 *Left*: stable solution (using the large time step $k = 0.01$) computed with the shifted Crank-Nicolson and the Fractional-Step-θ scheme. Recall the blow-up of the unstabilized Crank-Nicolson scheme in this case. *Right*: using the smaller time step $k = 0.001$ yields stable solutions for any time-stepping scheme. The unit of the time axis is s, whereas the drag unit is $\mathrm{kg/m\,s^2}$

References

1. Bungartz, H.J., Schäfer, M.: Fluid-Structure Interaction: Modelling, Simulation, Optimization, *Lecture Notes in Computational Science and Engineering*, vol. 53. Springer (2006)
2. Farhat, C., Geuzaine, P., Grandmont, C.: The discrete geometrical conservation law and the nonlinear stability of the ALE schemes for the solution of flow problems on moving grids. J. Comp. Phys. **174**, 669–694 (2001)
3. Fernández, M., Gerbeau, J.F.: Algorithms for fluid-structure interaction problems, pp. 307–346. Vol. 1 of Formaggia et al. [6] (2009)
4. Formaggia, L., Nobile, F.: A stability analysis for the arbitrary Lagrangian Eulerian formulation with finite elements. East-West Journal of Numerical Mathematics **7**, 105–132 (1999)
5. Formaggia, L., Nobile, F.: Stability analysis of second-order time accurate schemes for ALE-FEM. Comp. Methods Appl. Mech. Engrg. **193**(39–41), 4097–4116 (2004)
6. Formaggia, L., Quarteroni, A., Veneziani, A.: Cardiovascular Mathematics: Modeling and simulation of the circulatory system. Springer-Verlag, Italia, Milano (2009)
7. Heywood, J.G., Rannacher, R.: Finite-element approximation of the nonstationary Navier-Stokes problem part iv: Error analysis for second-order time discretization. SIAM Journal on Numerical Analysis **27**(2), 353–384 (1990)
8. Hron, J., Turek, S.: Proposal for numerical benchmarking of fluid-structure interaction between an elastic object and laminar incompressible flow, vol. 53, pp. 146–170. Springer-Verlag (2006)
9. Lesoinne, M., Farhat, C.: Geometric conservation laws for flow problems with moving boundaries and deformable meshes and their impact on aeroelastic computations. Comp. Methods Appl. Mech. Engrg. **34** (1996)
10. Luskin, M., Rannacher, R.: On the soothing property of the Crank-Nicolson scheme. Applicable Analysis **14**(2), 117–135 (1980)
11. Nobile, F.: Numerical approximation of fluid-structure interaction problems with applications to haemodynamics. Ph.D. thesis, École Polytechnique Fédérale de Lausanne (2001)
12. Rannacher, R.: On the stabilization of the Crank-Nicolson scheme for long time calculations (1986). Preprint
13. Rannacher, R.: Differences between the secant and the tangent Crank-Nicolson scheme. Personal Correspondance (2011)
14. Turek, S.: Efficient solvers for incompressible flow problems. Springer-Verlag (1999)
15. Wick, T.: Adaptive Finite Element Simulation of Fluid-Structure Interaction with Application to Heart-Valve Dynamics. Ph.D. thesis, University of Heidelberg (2011)
16. Wick, T.: Fluid-structure interactions using different mesh motion techniques. Comput. Struct. **89**(13–14), 1456–1467 (2011)

Various Flow Equations to Model the New Soil Improvement Method Biogrout

W.K. van Wijngaarden, F.J. Vermolen, G.A.M. van Meurs, and C. Vuik

Abstract We consider a mathematical model for Biogrout, which is a novel soil reinforcement technique based on Microbially Induced Carbonated Precipitation. We focus on an adaptation of the flow equation such that mass is conserved instead of volume. The adaptation is validated by a mass balance. Some numerical simulations are presented and used for the discussion on the various adjustments of the flow equation.

1 Introduction

Biogrout is a new soil reinforcement method that is based on Microbially Induced Carbonate Precipitation (MICP) [4]. Bacteria are placed and the reactants, urea ($CO(NH_2)_2$) and calcium chloride ($CaCl_2$), are flushed through the soil. The bacteria provide the hydrolysis of urea, and carbonate (CO_3^{2-}) is formed. Ammonium (NH_4^+) is formed as a side-product. In the presence of calcium (Ca^{2+}), the carbonate precipitates as calcium carbonate ($CaCO_3$). In [4], the biochemical reaction equations for the hydrolysis and precipitation are given. Combining these reactions, gives the overall Biogrout reaction equation:

$$CO(NH_2)_2(aq) + Ca^{2+}(aq) + 2H_2O(l) \rightarrow 2NH_4^+(aq) + CaCO_3(s). \qquad (1)$$

W.K. van Wijngaarden (✉) · G.A.M. van Meurs
Deltares, unit Geo Engineering, Postal Office 177, 2600 MH, Delft, The Netherlands
e-mail: Miranda.vanWijngaarden@Deltares.nl; Gerard.vanMeurs@Deltares.nl

F.J. Vermolen · C. Vuik
Delft Institute of Applied Mathematics, Delft University of Technology, Mekelweg 4, 2628 CD, Delft, The Netherlands
e-mail: F.J.Vermolen@tudelft.nl; C.Vuik@tudelft.nl

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_67,

The solid calcium carbonate forms bridges between the sand grains, which cause an increase in strength and stiffness of the soil. Because of its reinforcing properties, Biogrout can be used for, among others, the stabilization of soil, prior to tunnelling and for prevention of liquefaction of the subsoil resulting from earthquakes.

In [3], a mathematical model was constructed to describe the process by including equations for concentrations, flow and porosity. These equations are presented in Sect. 2. The partial differential equation for the flow, as displayed in Sect. 2, is based on the assumption that the volume of the fluid is conserved. This is verified by a mass balance calculation, which can also be found in Sect. 2. It appears that the volume of the fluid is not conserved. Two other partial differential equations are derived, which are based on the conservation of mass.

In Sect. 3, the numerical methods that are used to solve the system of equations are presented, as well as the simulations that have been done. In Sect. 4, the results of the application of the various partial differential equations for the flow are compared. In Sect. 5, some discussion and conclusions can be found.

2 The Mathematical Model

In this section, the model equations for the Biogrout process are presented and shortly discussed. The derivation can be found in [3].

The concentrations of the aqueous species are modelled through an advection-dispersion-reaction equation, see [1]:

$$\frac{\partial(\theta C^i)}{\partial t} = \nabla \cdot (\theta \mathsf{D} \cdot \nabla C^i) - \nabla \cdot (\mathbf{q} C^i) + n_i \theta r. \tag{2}$$

In this equation, θ is the porosity, C^i is the dissolved concentration of species i with M(=kmol/m^3) as a unit, D is the dispersion tensor, $\mathbf{v}$ is the pore water velocity, n_i is a constant that deals with the stoichiometry in the biochemical reaction Eq. (1) and r is the reaction rate of the production of calcium carbonate, which is a non-linear function of the urea concentration. Since the relevant aqueous species in the Biogrout process are urea, calcium and ammonium, we have $i \in \{\text{urea}, Ca^{2+}, NH_4^+\}$. From the stoichiometry of reaction (1), the values of n_i for the various aqueous species are given by: $n_{\text{urea}} = -1$, $n_{Ca^{2+}} = -1$, and $n_{NH_4^+} = 2$.

The left-hand side of Eq. (2) represents the accumulation. The first term at the right-hand side stands for the effect of dispersion and diffusion, the second term represents advection and the last term models the biochemical reaction.

For the reaction rate r of Eq. (1), a Monod equation has been used:

$$r = v_{max} \frac{C^{\text{urea}}}{K_m + C^{\text{urea}}}. \tag{3}$$

Here, v_{max} is the reaction rate constant and K_m is the saturation constant.

Table 1 Molar mass of urea, calcium chloride, ammonium chloride and water

$m_{urea} = 60.0551$ kg/kmol	$m_{CaCl_2} = 110.9840$ kg/kmol	$m_{NH_4Cl} = 53.4913$ kg/kmol
$m_{CaCO_3} = 100.0869$ kg/kmol	$m_{H_2O} = 18.0152$ kg/kmol	

Since it has been assumed that the non-aqueous calcium carbonate is not transported, there are no transport terms in the corresponding differential equation. The concentration of calcium carbonate $C^{\mathrm{CaCO_3}}$ can be calculated from the following differential equation:

$$\frac{\partial C^{\mathrm{CaCO_3}}}{\partial t} = m_{\mathrm{CaCO_3}} \theta r. \tag{4}$$

In this equation, $m_{\mathrm{CaCO_3}}$ is the molar mass of calcium carbonate, which is used to convert moles into mass.

Due to the precipitation of calcium carbonate, the porosity decreases. The following relation exists between the porosity and the calcium carbonate concentration:

$$\theta(t) = \theta(0) - \frac{C^{\mathrm{CaCO_3}}(t) - C^{\mathrm{CaCO_3}}(0)}{\rho_{\mathrm{CaCO_3}}}. \tag{5}$$

Here $\rho_{\mathrm{CaCO_3}}$ denotes the density of calcium carbonate.

It has been assumed that reaction (1) has no influence on the total volume of the fluid over the entire domain of computation and that the fluid is incompressible. This implies that the total volume of the fluid is conserved. Hence based on this hypothesis, the following partial differential equation was derived for the Darcy flow velocity $\mathbf{q}$:

$$\nabla \cdot \mathbf{q} = \frac{m_{\mathrm{CaCO_3}}}{\rho_{\mathrm{CaCO_3}}} \theta r. \tag{6}$$

Since the porosity decreases, due to the calcium carbonate precipitation, there is less space available for the fluid. This lack of space explains the non-zero right-hand side of Eq. (6). The last differential equation completes the set of equations that is necessary to simulate the Biogrout process.

Differential equation (6) is based on the assumption that the total volume of the fluid is conserved. With a mass balance calculation it is verified, whether this assumption is true.

As a relation for the density of the solution, we use

$$\rho = 1000 + 15.4996 C^{urea} + 86.7338 C^{Ca^{2+}} + 15.8991 C^{NH_4^+}, \tag{7}$$

as derived in [3]. Table 1 contains the molar mass of urea, calcium chloride ($CaCl_2$), ammonium chloride (NH_4Cl) and water (H_2O).

We consider $1\,\mathrm{m}^3$ of a 1 M urea/calcium chloride solution. All the urea and calcium chloride are converted to calcium carbonate and ammonium chloride, which

Table 2 The mass and mole balances of the reaction

	Initial state		Conversion		Final state	
	kmol	kg	kmol	kg	kmol	kg
Urea	1	60.0551	−1	− 60.0551	0	0
$CaCl_2$	1	110.9840	−1	−110.9840	0	0
NH_4Cl	0	0	+2	+106.9826	2	106.9826
$CaCO_3$	0	0	+1	+100.0869	(1)	(100.0869)
H_2O	51.6894	931.1943	−2	− 36.0304	49.6894	895.1645

means that 1 kmol urea, 1 kmol calcium chloride and 2 kmol water disappear from the solution and 2 kmol ammonium chloride and 1 kmol calcium carbonate are formed. The calcium carbonate precipitates out of the solution. The ammonium chloride stays in the solution.

Table 2 displays the mass and mole balances of the reaction.

We define $\tilde{V}$ as the volume after conversion. The density of the solution equals $\rho = 1,000 + 15.8991 \cdot \frac{2}{\tilde{V}}$ and the mass of the liquid equals $1,000\,\tilde{V} + 15.8991 \cdot 2$. From the mass balance in Table 2, it follows that this must be equal to 1002.1471 kg. Therefore, $\tilde{V} = 0.97035\,\mathrm{m}^3$, which is not equal to $1\,\mathrm{m}^3$.

From this mass balance calculation, it follows that the hydrolysis of urea and the precipitation of calcium carbonate do influence the volume of the fluid and that the assumption of conservation of fluid volume is not valid. Hence, the differential equation for the flow, based on this assumption, should be adapted.

Therefore, two alternative partial differential equations are introduced. The first one is almost similar to the previously used differential equation (6), but corrects for the shrinking liquid volume. In the previous paragraphs, we calculated that for each converted kmol urea and calcium chloride, the total liquid volume shrinks with $1 - \tilde{V} = 0.02965\,\mathrm{m}^3$. Therefore, an extra term is added to Eq. (6) to correct for this phenomenon. The following alternative partial differential equation for the flow is derived:

$$\nabla \cdot \mathbf{q} = \left(\frac{m_{CaCO_3}}{\rho_{CaCO_3}} - (1 - \tilde{V}) \right) \theta r. \tag{8}$$

As a second alternative partial differential equation for the flow, the following equation was derived from a more physical point of view:

$$\frac{\partial \left(\rho \theta \right)}{\partial t} = -\nabla \cdot \left(\rho \mathbf{q} \right) - m_{CaCO_3} \theta r. \tag{9}$$

The left-hand side represents the accumulation of mass in the pores. The first term at the right-hand side models mass flow and the last term stands for the mass that disappears from the fluid, as a result of precipitation.

Table 3 The process variables that can be varied. In each simulation set, only one variable is varied, while the **bold values** are assigned to the other variables

Process variable	Value					
q_{in} [m/h]	0	0.001	**0.01**	0.1	1	
θ_0 [1]		0.1	**0.3**	0.5		
v_{max} [kmol/m^3/h]		0.0036	**0.036**	0.36		
c_{in} [kg/m^3]		0.5	**1**	2	3	4

3 Strategy and Numerical Methods

The aim of this research is to find out whether an alternative differential equation for the flow will result into a different calcium carbonate content, between realistic bounds for the process variables that can be chosen freely. In order to do so, these process variables are varied. As an experimental set-up, we take a one dimensional configuration, which corresponds to a column experiment. The inflow is at the left-hand side and the outflow at the right-hand side.

The process variables that can be chosen are: the inflow velocity q_{in}, the inflow concentration of urea and calcium chloride c_{in} and the maximal bacterial activity v_{max}. The latter can be chosen during the cultivation or by diluting a suspension with a high activity. The (initial) porosity θ_0 is given, but varies initially locally. In laboratory experiments one can more or less adapt the porosity by filling the experimental set-up with sand.

Several computer simulations have been done to examine whether the various differential equations for the flow lead to different calcium carbonate contents. These computations have been done as follows: Certain values have been assigned to the four process variables. These values are the bold values that can be found in Table 3, which forms the basis combination. Then, four sets of simulations are done. In each set, only one variable from this basis combination is adjusted. The values are given in Table 3, again. The results from this comparison can be found in Sect. 4.

Initially, all concentrations are zero. Only for the case $q_{in} = 0$, we need to take an alternative initial condition for the concentration of urea and calcium chloride, otherwise nothing will happen. As an initial condition for the urea and calcium chloride concentration, we take:

$$C^i(t,0) = \begin{cases} 1 \text{ for } 0 \le x \le 0.90; \\ 0 \text{ for } 0.90 < x \le 1, \end{cases} \tag{10}$$

with $i \in \{\text{urea}, Ca^{2+}\}$. The initial porosity is given in Table 3.

As a boundary condition for the flow at the inflow boundary we take the Dirichlet boundary condition $q(0,t) = q_{in}$. The values of q_{in} can be found in Table 3. For the concentration of urea and calcium chloride, we take the following Robin boundary condition at the inflow boundary: $(D\theta\nabla c - qc) \cdot n = q_{in}c_{in}$, which implies that the amount of urea and calcium chloride that enters the domain per unit of time and surface equals $q_{in}c_{in}$. Again, the values of c_{in} are given in Table 3. Since no ammonium chloride is injected, we take $(D\theta\nabla c - qc) \cdot n = 0$ as a boundary

condition for the concentration of ammonium chloride at the inflow boundary. At the outflow boundary we choose the homogeneous Neumann condition $D\theta\nabla c = 0$ for the concentration of all aqueous species, which corresponds to an advective flux. The simulated time is 100 h, unless stated otherwise.

We use the Standard Galerkin Finite Element Method to solve the model equations. For more information on the numerical methods we refer to [2] and [3], where this has been reported in more detail, also for the higher-dimensional cases.

4 Results

In this section some results (Fig. 1) are shown from the comparison between the two alternative partial differential equations for the flow, (8) and (9), and the previously used differential equation (6). The left graphs represent the Darcy flow velocity and the right ones the calcium carbonate concentration.

The top graphs of Fig. 1 show the Darcy flow velocity and the calcium carbonate concentration as a function of location at time t = 100 h. The values that have been assigned to the process variables are the bold values in Table 3. As can be seen, the flow that is calculated from the old differential equation differs only by 3 % from the flow that is calculated from the alternatives. Hence, this adaptation only has a minor effect on the calcium carbonate concentration. The results from the variation of θ_0, v_{max}, c_{in} and q_{in} are similar: the calculated flows show a small difference, the calcium carbonate content is very similar.

The variation of θ_0, v_{max} and (non-zero) q_{in} results into a maximal difference in calcium carbonate content of at most 2 kg/m^3, which corresponds to a relative difference in the order of 5 %.

The difference in calcium carbonate content increases for an increasing c_{in}. For $c_{in} = 4$ M, the maximal difference in calcium carbonate concentration is 5 kg/m^3 for the first alternative differential equation and 3 kg/m^3 for the second one. This difference is still in the order of the measurement error. However, since the solubility of urea in water is 18 M and the solubility of calcium chloride in water is 7 M, one might wonder whether higher urea/calcium chloride concentrations will result into larger differences in calcium carbonate. However, since a concentration of 4 M is already toxic for bacteria, high concentrations will never be used in the Biogrout process. The middle graphs of Fig. 1 show the flow and the calcium carbonate content for $c_{in} = 4$ M.

The bottom graphs of Fig. 1 show the Darcy flow velocity and the calcium carbonate concentration at time $t = 25$ h for the zero inflow velocity case. In this case the simulated time is 25 h, since at time t = 100 h all the urea and calcium chloride have reacted, so there is no driving force for a flow any more. Note that the difference in Darcy flow velocity is more pronounced in the bottom left graph than in the other graphs. The initial conditions for the urea and calcium chloride concentration, as given in Eq. (10), have been chosen in such a way, that the difference in calcium carbonate content is really large. This difference is 20 kg/m^3

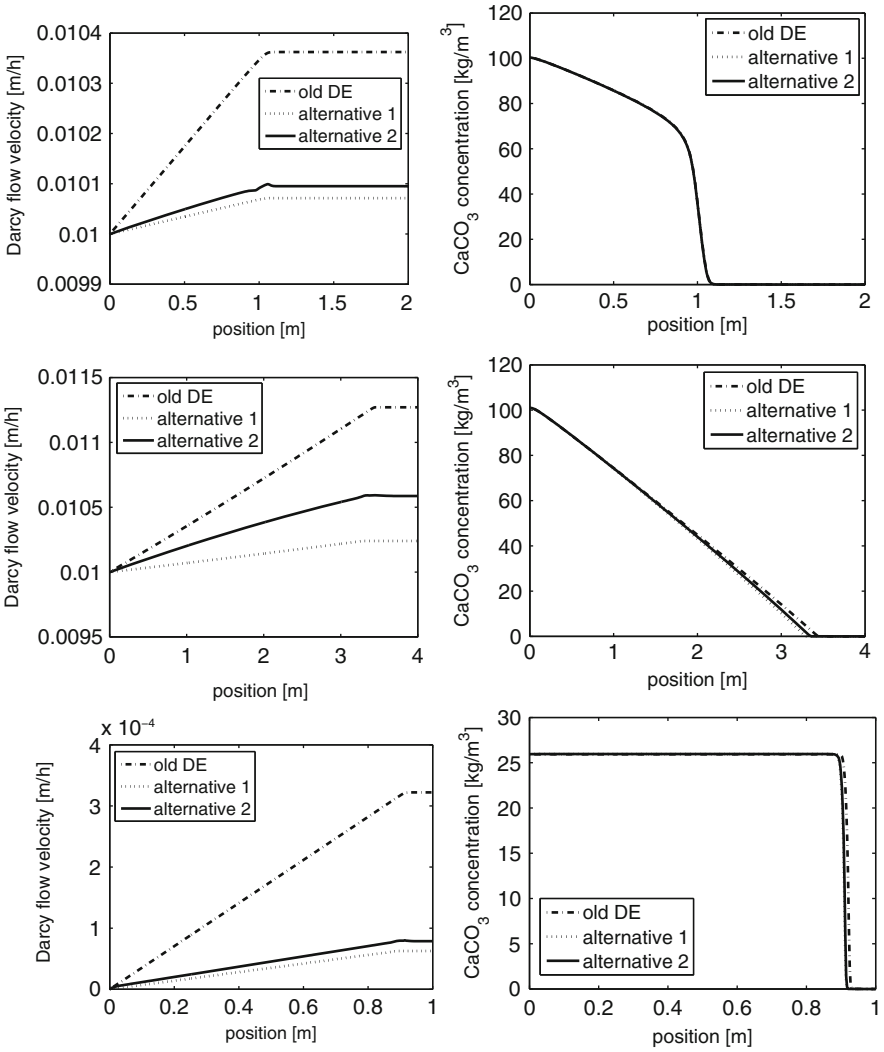


Fig. 1 The Darcy flow velocity (*left graphs*) and the calcium carbonate concentration (*right graphs*) as a function of location, for the bold values from Table 3 at time t = 100 h (*top graphs*), for inflow velocity $c_{in} = 4$ M at time t = 100 h (*middle graphs*) and for $q_{in} = 0$ m/h at time t = 25 h (*bottom graphs*)

for the first alternative differential equation and 19 kg/m^3 for the second one. The graphs are very similar, however. Both graphs have a constant calcium carbonate content for the first 0.9 m, followed by a steep front. The difference in the location of this steep front is in the order of only 1 cm. Due to the very steep front, this difference in location results into a large difference in calcium carbonate content. The steep front in calcium carbonate content results from the steep front in urea

and calcium chloride that has been chosen as an initial condition. In practice, the gradient will be much smaller due to dispersion and diffusion.

5 Discussion and Conclusions

From the results in Sect. 4, we conclude that the three different partial differential equations for the flow lead to different flow patterns. However, the graphs of the corresponding calcium carbonate content are very similar. In most cases the maximal difference between the calcium carbonate content is at most $2\,\mathrm{kg/m^3}$, which corresponds to a relative error in the order of 5 %.

From the variation of the inflow concentration of urea and calcium chloride c_{in}, it is concluded that a larger value of c_{in} leads to a higher maximal difference in calcium carbonate content. But since the inflow concentration that is used is limited, due to its toxicity for the bacteria, the error stays small for realistic values of c_{in}.

The results of the zero inflow velocity case show that steep gradients in the urea and calcium chloride concentration lead to steep gradients in the calcium carbonate content. A small difference in flow then leads to a small difference in the position of the front (in the order of 1 cm) and to a high difference in calcium carbonate content in that small region. However, since the calcium carbonate fronts are really close and such steep gradients in urea and calcium chloride are not likely to occur, this case does not lead to any important differences in calcium carbonate content.

Therefore, we conclude that the choice of the differential equation for the flow hardly influences the calcium carbonate content for realistic values of the process variables. Since the process variables in [2] and [3] are within the ranges specified in Table 3, the results in both articles are still valid.

Although the results are very similar, we will no longer use the previous differential equation for the flow. Instead, we will use one of the alternative differential equations, since they do not violate the requirement of conservation of mass. We choose the first alternative differential equation (8) to use from now on, since the first alternative is simpler and more stable than the second one.

We realize that the alternative differential equations, derived in Sect. 2 are based, among others, on the empirical relation between the density and the various concentrations (7).

References

1. Zheng, C., Bennett, G D. : Applied Contaminant Transport Modeling, Van Nostrand Reinhold, New York (1995)
2. Van Wijngaarden, W.K., Vermolen, F.J., van Meurs, G.A.M., Vuik, C.: Modelling the New Soil Improvement Method Biogrout: Extension to 3D. In: G. Kreiss et al. (eds.) Numerical Mathematics and Advanced Applications, 893–900 (2009)

3. Van Wijngaarden, W.K., Vermolen, F.J., van Meurs, G.A.M., Vuik, C.: Modelling Biogrout: A New Ground Improvement Method Based on Microbial-Induced Carbonate Precipitation. Transp. Porous Med. **87**, 397–420 (2011)
4. Whiffin, Victoria S., van Paassen, Leon A. and Harkes, Marien P.: *Microbial Carbonate Precipitation as a Soil Improvement Technique*, Geomicrobiology Journal, *24:5*, (2007), 417–423.

A Fast GPU-Accelerated Mixed-Precision Strategy for Fully Nonlinear Water Wave Computations

S.L. Glimberg, A.P. Engsig-Karup, and M.G. Madsen

Abstract We present performance results of a mixed-precision strategy developed to improve a recently developed massively parallel GPU-accelerated tool for fast and scalable simulation of unsteady fully nonlinear free surface water waves over uneven depths (Engsig-Karup et al., Int J Num Meth, 2011). The underlying wave model is based on a potential flow formulation, which requires efficient solution of a Laplace problem at large-scales. We report recent results on a new mixed-precision strategy for efficient iterative high-order accurate and scalable solution of the Laplace problem using a multigrid-preconditioned defect correction method. The improved strategy improves the performance by exploiting architectural features of modern GPUs for mixed precision computations and is tested in a recently developed generic library for fast prototyping of PDE solvers. The new wave tool is applicable to solve and analyze large-scale wave problems in coastal and offshore engineering.

1 Introduction

Recent development significantly improves the strategy proposed by Li and Fleming in [9] to simulate fully nonlinear water waves. A flexible-order finite difference algorithm for solving the governing equations in two (Bingham and Zhang [4]) and three space dimensions (Engsig-Karup et al. [6]) enables efficient, scalable and low-storage solution of the equations. Recent developments in modern many-core hardware and programming tools for general-purpose scientific computing, suggest that a combination could further improve the overall performance.

S.L. Glimberg (✉) · A.P. Engsig-Karup · M.G. Madsen
Department of Informatics and Mathematical Modelling, Technical University of Denmark, 2800 Kgs Lyngby, Denmark
e-mail: slgl@imm.dtu.dk; apek@imm.dtu.dk; morten.gorm.madsen@gmail.com

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_68,

In recent work [7], we have demonstrated that it is now possible to significantly reduce the barriers for practical use of full potential flow theory as the modeling basis for efficient solution of coastal and offshore engineering problems. Our strategy was to do proof-of-concept by utilizing modern Graphics Processing Units (GPUs) for massively parallel computations using a heterogeneous CPU-GPU hardware setup. Interestingly, such a hardware setup constitutes what can be considered an affordable standard consumer desktop environment.

To establish the model as an efficient massively parallel tool we have both redesigned and reimplemented the entire algorithm using a newly developed library for PDE solver proto-typing. The library enables efficient utilization of allocated hardware resources, targeting modern many-core GPUs. Algorithmic efficiency is achieved by solving the computational bottleneck problem iteratively with a defect correction method, preconditioned by a robust multigrid method. This strategy gives more than one order of magnitude in both problem size and practical speedup (relative to optimized single-threaded CPU code).

1.1 Governing Equations

We present recent progress on the development of the OceanWave3D model [4, 6]. In short, the flexible-order finite difference OceanWave3D model is based on a unified potential flow formulation. These model equations can account for fully nonlinear and dispersive waves within the breaking limit and under the assumption of irrotational inviscid flow. The temporal derivatives for the surface variables, i.e. the free surface elevation η and the velocity potential $\tilde{\phi}$ is given by

$$\partial_t \eta = -\nabla\eta \cdot \nabla\tilde{\phi} + \tilde{\omega}(1 + \nabla\eta \cdot \nabla\eta) \tag{1}$$

$$\partial_t \tilde{\phi} = -g\eta - \frac{1}{2}(\nabla\tilde{\phi} \cdot \nabla\tilde{\phi} - \tilde{\omega}^2(1 + \nabla\eta \cdot \nabla\eta)), \tag{2}$$

where $\nabla = [\partial_x\ \partial_y]^T$, $\tilde{\omega} = \partial_z\phi|_{z=\eta}$ and g is the gravitational acceleration. In order to integrate these equations in time, the vertical velocity on the surface $\tilde{\omega}$, must be determined from the full potential inside the domain. The following Laplace equation along with boundary conditions uniquely defines the full velocity potential

$$\begin{aligned} \phi &= \tilde{\phi}, \quad && z = \eta \\ \nabla^2\phi + \partial_{zz}\phi &= 0, && -h \le z < \eta \\ \partial_z\phi + \nabla h \cdot \nabla\phi &= 0, && z = -h, \end{aligned} \tag{3}$$

Where h is the still water depth. Notice that the Laplace problem is of three dimensions, whereas the surface time integration is only of two dimensions. Thus, the computational effort to solve the discretized Laplace problem (3) is the most time

consuming part of a numerical solver for this problem. In the following we focus on the numerical approach to solve (3) efficiently on many-core GPUs. In practice we actually solve the so-called σ-transformed version of (3), in order to avoid time changing domains and variable finite difference coefficients from approximating the derivatives. See [9] or [6] for details on the transformed equations.

2 Development of a Massively Parallel Wave Analysis Tool

The flexible-order finite difference scheme presented by Engsig-Karup et al. [6] was originally implemented as a stand-alone serial code. The tool was referred to as OceanWave3D. In a recent proof-of-concept study the algorithmic strategy for the OceanWave3D model was first improved and then a massively parallel implementation was carried out and tested on a single GPU [7] with significant performance improvements. The flexible-order finite difference operators was implemented as matrix-free compact stencil operators, in order to further minimize the memory overhead of storing identical entries and avoiding the extra index tables required by traditional sparse matrix formats. Figure 1 is replicated from [7] and illustrates linear scalability of absolute timings as the problem size increases along with speedups relative to optimized single-threaded CPU code. Recently, a library for high-performance PDE solver proto-typing has been established and the OceanWave3D strategy was again transferred to this new library. The existing dedicated GPU implementation has been used as a reference, to ensure no significant performance loss using the new high-level library. A short outline of the library is presented next.

2.1 A Library for Fast PDE Solver Proto-Typing

Our generic high-performance C++ library is subject to ongoing development and improvements within our research group. The purpose is to enable fast proto-typing of efficient massively parallel solvers, inspired by the PETSc toolkit library [2]. Our library facilitates massively parallelization through GPU computing and contains components for various iterative strategies for solution of large linear systems. The goal has been to create a portable and reusable framework without losing noticeable performance – a common tradeoff between generality and dedicated solvers.

The generic nature of our library enables the end users to easily change solver parts through type bindings. The backbone of the library is a generic vector class. It takes two template parameters to define the container type along with a memory space identifier, inspired by the Thrust and Cusp GPU libraries [3,8]. The following simple example illustrates how to set a vector type definition such that the program uses the GPU for memory storage and computations.

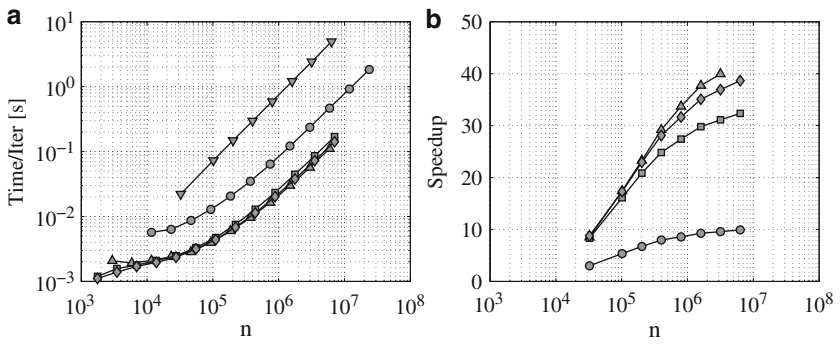


Fig. 1 Scalability tests and performance comparisons in double precision arithmetic for Quadro FX 5800 (−•−), GeForce GTX 480 (−▲−), C2050 with ECC (−■−) and C2050 without ECC (−♦−) versus CPU (single thread) code (−▼−). Sixth order spatial discretization employed. The iterative defect correction method has been left-preconditioned with a Zebra Line Gauss-Seidel V-cycle multigrid strategy on each architecture. (**a**) Absolute timings. (**b**) Speedup relative to CPU (single thread)

```
// Make a type definition to determine the vector type of the coming program
typedef vector<float,device_memory> vector_type;
vector_type x(100);      // Create vector x in GPU memory
vector_type y(100);      // Create vector y in GPU memory
y.axpy(2.f,x);           // Calc. y = a*x + y on the GPU
```

The above example might seem trivial, but the use of type definitions can be taken further, using so called type binders. Setting up our free surface solver looks similar to the following code example, using the predefined type binder class potential_flow_solver_types.

```
// Potential flow setup
typedef free_surface::potential_flow_solver_types<
        vector_type                                 // Vector object
      , solvers::multigrid<multigrid_types>         // Laplace solver
      , integration::ERK4                           // Time integrator
> potential_flow_types;
```

Afterwards, the solver object is instantiated with this type binder definition given as template argument. The solver hereby implicitly knows all necessary types to use within its own implementation. Consequently, parts can be treated as building blocks to make up the entire solver. If for example the user wants to use another time integrator or Laplace solver, the corresponding line is exchanged with an alternative implementation, either user specific or from the library itself. Notice that the multigrid solver is a template class itself that depends on another type binder, also specified by the user. Concepts of template based programming is well presented in the book by Vandevoorde and Josuttis [11].

The Laplace equation (3) is solved with an iterative multigrid-preconditioned defect correction method, a variation of the iterative refinement method [10]. In practice, the defect correction method turns out to be as effective as a reference GMRES solver. Furthermore the defect correction method has two important prop-

Algorithm 1 DC method for approximate solution of $Ax = b$

1	Choose $x^{[0]}$	/* initial guess */
2	$k = 0$	
3	**Repeat**	
4	$r^{[k]} = b - Ax^{[k]}$	/* high order defect */
5	Solve $M\delta^{[k]} = r^{[k]}$	/* preconditioner */
6	$x^{[k+1]} = x^{[k]} + \delta^{[k]}$	/* defect correction */
7	$k = k + 1$	
8	**Until** convergence or $k > k_{max}$	

erties: (i) Constant minimal memory footprint. (ii) Few synchronization barriers. These two properties make it very attractive from a parallel point of view. A textbook recipe of the defect correction method is given in Algorithm 1. This algorithm is implemented into our library in the same generic way as previously described. Building the solver using a predefined type binder class could look as follows, assuming that proper types for the vector, matrix, and preconditioner are set beforehand.

```
1  typedef solvers::defect_correction_types<
2        vector_type
3      , matrix_type
4      , preconditioner_type> dc_types;   // DC type binder
5  typedef solvers::defect_correction<dc_types> dc_solver_type;
6
7  // Create solver, assume vectors (x,b) and matrices (A,P) are already created
8  dc_solver_type solver( A );            // Create solver
9  solver.set_preconditioner( P );        // Set preconditioner
10 solver.solve(x, b);                    // Solve Ax = b
```

From building blocks in the library, we have set up a 2D time integration solver for the fully nonlinear free surface waves. The library has tools for most of the needed components, such as the time integration scheme, solver for the linear system, printing functionality and so on. The main functionality that the user has to deliver, is an implementation of the matrix-vector product from the discretization of (3), required to calculate the residual in line 4 of Algorithm 1. Algorithmic efficiency is achieved with a multigrid preconditioning strategy based on a low-order discretization of the linearized system matrix (see [6]) and red-black Gauss-Seidel smoothening. This smoother must also be made available to the multigrid solver by the user.

2.2 *Improving Defect Correction with Mixed Precision*

In order to further improve the nonlinear free surface solver, a mixed precision strategy has recently been added to the defect correction scheme. The purpose of the mixed precision algorithm is to reduce the overall computational and storage requirements by introducing low (single) precision arithmetics.

The advantage from a memory perspective is obvious: single precision numbers take up only half the storage of a double precision number (32 bits vs. 64 bits). Thus, storage and bandwidth requirements are halved. The computational demands are also reduced. However, this is somewhat more hardware dependent. Most modern CPU architectures obtain twice the performance for single precision execution compared to double precision, see [5]. On GPU architectures this relation might be more distinct. On a TESLA S1070 computing system, single precision operations are up to 12 times faster.

As noted in [1], any refinement process is a candidate to benefit from mixed precision computations, since often only the refinement itself needs to be in double precision arithmetic. Rewriting the defect correction scheme from Algorithm 1 into a single expression for iterative refinement of x at iteration $k+1$ gives

$$x^{[k+1]} = x^{[k]} + M^{-1}(b - Ax^{[k]}). \tag{4}$$

Assuming that the iterative scheme converges towards the exact solution, the correction term $M^{-1}(b - Ax^{[k]})$ reduces in magnitude for each iteration until an acceptable accuracy threshold can be met. If both $x^{[k]}$ and the correction term are in single precision, round off errors naturally occur earlier than they would in double precision. The trick is to calculate only the correction term in single precision and do the update in double precision. Since the correction term is approaching zero, the values are well represented in single precision and the double precision update only suffer from rounding errors when the correction approaches values near $\sim 10^{-16}$. Thus, we get a double precision accurate solution, while being able to do parts of the calculations in single precision. Applying this technique to the defect correction scheme, the preconditioning step in line 5 of Algorithm 1, is simply executed in pure single precision arithmetics.

With this strategy we have been able to further improve the OceanWave3D model. Performance results for the mixed precision strategy on a Tesla C2050 are given in Fig. 2. The C2050 has a 2 : 1 ratio on the peak performance for double precision vs. single precision. However, the algorithm is memory bound, so we expect the observed behavior to be caused by the 2 : 1 restriction on the memory bandwidth. As expected, a pure single precision iteration takes approximately half the time (x1.9 faster) for larger systems. The mixed precision strategy is however the only one that would give a high precision solution and therefore the only fair comparison to the double precision strategy. Roughly a speedup of x1.6 is achieved for large enough systems. Absolute timings and relative speedups of the Laplace solver are depicted in Fig. 3. The double precision timings are slightly better than the ones previously presented in Fig. 1 from [7]. This is not surprisingly since the 3D finite difference operations in [7] are more expensive than the 2D operations in the present work. Still, we would expect an extension to 3D of the present solver to give results in the same range as the dedicated 3D solver. Taking also the mixed precision extension into consideration, we expect a 3D solver to gain about the same x1.6 extra speedup as well.

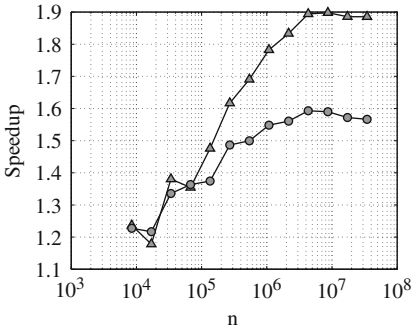


Fig. 2 Speedups for a defect correction iteration using single precision (−▼−) and mixed precision (− • −) relative to double precision. All timings are on a Tesla C2050 GPU

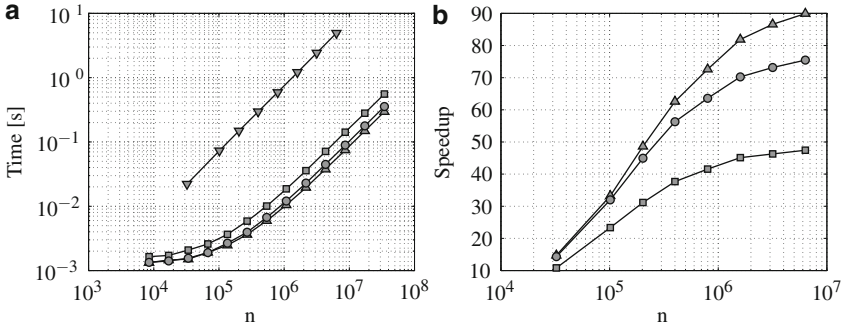


Fig. 3 Scalability tests and performance comparisons on Tesla C2050 in single precision (−▲−), double precision (−■−), mixed precision (− • −), and CPU (single thread) code (−▼−). Sixth order spatial discretization employed. The iterative defect correction method has been left-preconditioned with a Gauss-Seidel V-cycle multigrid strategy on each architecture. (**a**) Absolute timings. (**b**) Speedup relative to CPU (single thread)

3 Concluding Remarks

The potential flow equations describing fully nonlinear water waves have been efficiently solved and improved from previous work [7]. A highly generic GPU-based library has been developed, not only to solve the present equations, but also a broader range of PDEs that can be well discretized in a finite difference manner. The library is still at an early state and under continuously development. We expect that the library will ease future development of PDE solvers for a variety of physical problems and simulations. Results indicate that the library does not suffer from serious overhead, as performance results are comparable to an existing dedicated solver for the same model problem. Future work is to confirm that this indication is valid, by assembling a full 3D solver using library components.

Furthermore we illustrated how to easily extend the defect correction method in order to utilize a fast mixed precision strategy, by computing the preconditioning step in pure single precision arithmetics. This approach gives an additional x1.6 speedup on the Tesla C2050 GPU architecture. Combining these results we are approaching almost two orders of magnitude in relative speedup compared to the optimized single threaded CPU reference code from previous work [6].

Ongoing work is also concerned with large-scale modelling, in which the discretized equations does not fit into the memory of one GPU. A domain decomposition strategy is thus necessary to decompose memory across multiple GPUs. In this case MPI is used for the communication between nodes. The impact on performance of transferring artificial boundary information between nodes is to be investigated in future work.

References

1. Baboulin, M. and Buttarib, A. and Dongarra, J. and Kurzak, J. and Langouc, J. and Langou, J. and Luszczek, P. and Tomov, S.: Accelerating scientific computations with mixed precision algorithms. Comp. Phys. Comm. **180**, 2526–2533, (2009)
2. Balay, S. and Brown, J. and Buschelman, K. and Gropp, W. D. and Kaushik, D. and Knepley, M. G. and McInnes, L. C. and Smith, B. F. and Zhang, H.: PETSc, version 3.2. (2011) http://www.mcs.anl.gov/petsc
3. Bell, N. and Garland, M.: Cusp: Generic Parallel Algorithms for Sparse Matrix and Graph Computations, version 0.1.0. (2010) http://cusp-library.googlecode.com
4. Bingham, H. B. and Zhang, H.: On the accuracy of finite-difference solutions for nonlinear water waves. J. Engng. Math. **58**, 211–228, (2007)
5. Buttari A. and Dongarra, J. and Langou, J. and Langou, J. and Luszczek, P. and Kurzak J.: Mixed Precision Iterative Refinement Techniques for the Solution of Dense Linear Systems. Int. J. Hi. Perf. Comp. App. **21**, 457–466 (2007)
6. Engsig-Karup, A.P. and Bingham, H.B. and Lindberg, O.: An efficient flexible-order model for 3D nonlinear water waves. J. Comp. Phys. **228**, 2100–2118, (2009)
7. Engsig-Karup, A. P. and Madsen, M. G. and Glimberg, S. L.: A massively parallel GPU-accelerated model for analysis of fully nonlinear free surface waves. Int. J. Num. Meth. Fluids. (2011)
8. Hoberock, J. and Bell, N.: Thrust: A Parallel Template Library, version 1.3.0. (2010) http://www.meganewtons.com/
9. Li, B. and Fleming, C. A.: A three dimensional multigrid model for fully nonlinear water waves. Coast. Engng. **30**, 235–258, (1997)
10. Martin, R. S. and Peters, G. and Wilkinson J. H.: Handbook Series Linear Algebra Iterative Refinement of the Solution of a Positive Definite System of Equations. Num. Math. **8**, 203–216 (1966)
11. Vandevoorde, D. and Josuttis, N. M.: C++ Templates: The Complete Guide. Addison-Wesley Professional, (2002)

3D Helmholtz Krylov Solver Preconditioned by a Shifted Laplace Multigrid Method on Multi-GPUs

H. Knibbe, C.W. Oosterlee, and C. Vuik

Abstract We are focusing on an iterative solver for the three-dimensional Helmholtz equation on multi-GPU using CUDA (Compute Unified Device Architecture). The Helmholtz equation discretized by a second order finite difference scheme is solved with Bi-CGSTAB preconditioned by a shifted Laplace multigrid method. Two multi-GPU approaches are considered: data parallelism and split of the algorithm. Their implementations on multi-GPU architecture are compared to a multi-threaded CPU and single GPU implementation. The results show that the data parallel implementation is suffering from communication between GPUs and CPU, but is still a number of times faster compared to many-cores. The split of the algorithm across GPUs limits communication and delivers speedups comparable to a single GPU implementation.

1 Introduction

As it has been shown in paper [5] the implementation of numerical solvers for indefinite Helmholtz problems with spatially dependent wavenumber, such as Bi-CGSTAB and IDR(s) preconditioned by shifted Laplace multigrid method on a GPU is more than 25 times faster than on a single CPU. Comparison of single GPU to a single CPU is important but it is not representative for problems of realistic size.

H. Knibbe (✉) · C. Vuik
Delft University of Technology, Delft, Netherlands
e-mail: hknibbe@gmail.com; c.vuik@tudelft.nl

C.W. Oosterlee
Dutch National Research Centre for Mathematics and Computer Science (CWI), Delft University of Technology, Delft, Netherlands
e-mail: c.w.oosterlee@cwi.nl

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_69,

By realistic problem size we mean three-dimensional problems which lead after discretization to linear systems of equations with more than one million unknowns. Such problems arise when modeling a wavefield in geophysics.

Problems of realistic size are too large to fit in the memory of one GPU, even with the latest NVIDIA Fermi graphics card (see [6]). One solution is to use multiple GPUs. The currently widely used architecture consists of a multi-core connected to one or at most two GPUs. Moreover, in most of the cases those GPUs have different characteristics and memory size. A setup with four or more identical GPUs is rather uncommon, but it would be ideal from a memory point of view. It implies that the maximum memory is four times or more than on a single GPU. However GPUs are connected to a PCI bus and in some cases two GPUs share the same PCI bus, this creates data transfer limitation. To summarize, using multi-GPUs increases the total memory size but data transfer problems appear.

The aim of this paper is to consider different multi-GPU approaches and understand how data transfer affects performance of a Krylov solver with shifted Laplace multigrid preconditioner for the three-dimensional Helmholtz equation.

2 Helmholtz Equation and Solver

The Helmholtz equation in three dimensions for a wave problem in a heterogeneous medium is considered

$$-\frac{\partial^2\phi}{\partial x^2}-\frac{\partial^2\phi}{\partial y^2}-\frac{\partial^2\phi}{\partial z^2}-(1-\alpha i)k^2\phi=g, \tag{1}$$

where $\phi=\phi(x,y,z)$ is the wave pressure field, $k=k(x,y,z)$ is the wavenumber, $\alpha \ll 1$ is the damping coefficient, $g=g(x,y,z)$ is the source term. The corresponding differential operator has the form $\mathscr{A}=-\Delta-(1-\alpha i)k^2$, where Δ denotes the Laplace operator. The problem is given in a cubic domain $\Omega=[(0,0,0),(X,Y,Z)]$, $X,Y,Z\in\mathbb{R}$. A first order radiation boundary condition is applied $\left(-\frac{\partial}{\partial\eta}-ik\right)\phi=0$, where η is the outward normal vector to the boundary (see [2]). Discretizing equation (1) using the 7-point central finite difference scheme gives the following linear system of equations: $A\phi=g$, $A\in\mathbb{C}^{N\times N}$, $\phi,g\in\mathbb{C}^N$, where $N=n_xn_yn_z$ is a product of the number of discretization points in the $x-$, $y-$ and $z-$directions. Note that the closer the damping parameter α is set to zero, the more difficult it is to solve the Helmholtz equation. We are focusing on the original Helmholtz equation with $\alpha=0$.

As a solver for the discretized Helmholtz equation we have chosen the Bi-CGSTAB method preconditioned by shifted Laplace multigrid method with matrix-dependent transfer operations and a Gauss-Seidel smoother, (see [3]). It has been shown in [5] that this solver is parallelizable on CPUs as well as on a single GPU and provides good speed-up on parallel architectures. The prolongation in

this work is based on the three dimensional matrix-dependent prolongation for real-valued matrices described in [7]. This prolongation is also valid at the boundaries. The restriction is chosen as full weighting restriction. As a smoother the multi-colored Gauss-Seidel method has been used. In particular, for 3D problems the smoother uses eight colors, so that the color of a given point will be different from its neighbours.

Since our goal is to speed up the Helmholtz solver with the help of GPUs, we still would like to keep the double precision convergence rate of the Krylov method. Therefore Bi-CGSTAB is implemented in double precision. For the preconditioner, single precision is sufficient for CPU as well as GPU.

3 Multi-GPU Implementation

For our numerical experiments NVIDIA [6] provided a Westmere based 12-cores machine connected to 8 GPUs Tesla 2050 as shown on Fig. 1. The 12-core machine has 48 GB of RAM. Each socket has 6 CPU cores Intel(R) Xeon(R) CPU X5670 @ 2.93 GHz and is connected through 2 PCI-buses to 4 graphics cards. Note that two GPUs are sharing one PCI-bus connected to a socket. Each GPU consist of 448 cores with clock rate 1.5 GHz and has 3 GB of memory.

In the experiments CUDA version 3.2[1] is used. All experiments on CPU are done using a multi-threaded CPU implementation (pthreads).

In general GPU memory is much more limited than CPU memory so we chose a multi-GPU approach to be able to solve larger problems. The implementation on a single GPU of major components of the solver such as vector operations, matrix-vector-multiplication or the smoother has been described in [5]. In this section we focus on the multi-GPU implementation.

There are two ways to do computations on multi-GPU: push different Cuda contexts to different GPUs (see [6]) or create multiple threads on the CPU, where each thread communicates with one GPU. For our purposes we have chosen the second option, since it is easier to understand and implement.

Multiple open source libraries for multi-threading have been considered and tested. For our implementation of numerical methods on a GPU the main requirement for multi-threading was that a created thread stays alive to do further processing. It is crucial for performance that a thread remains alive as a GPU context is attached to it. Pthreads has been chosen as we have total control of the threads during the program execution.

[1] During the work on this paper, the newer version of CUDA 4.0 has been released. It was not possible to have the newer version installed on all systems for our experiments. That is why for consistency and comparability of experiments, we use the previous version

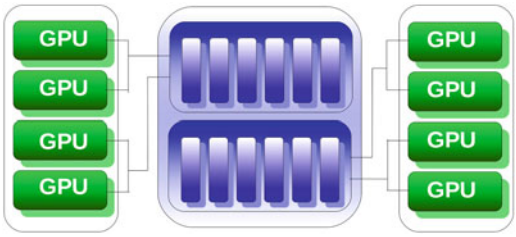


Fig. 1 NVIDIA machine with 12 Westmere CPUs and 8 Fermi GPUs, where two GPUs share a PCI bus connected to a socket

There are several approaches to deal with multi-GPU hardware:

1. *Domain-Decomposition approach*, where the original continuous or discrete problem is decomposed into parts which are executed on different GPUs and the overlapping information (halos) is exchanged by data transfer. This approach can however have difficulties with convergence for higher frequencies (see [4]).
2. *Data-parallel approach*, where all matrix-vector and vector-vector operations are split between multiple GPUs. The advantage of this approach is that it is relatively easy to implement. However, matrix-vector multiplication requires exchange of the data between different GPUs, that can lead to significant data transfer times if the computational part is small. The convergence of the solver is not affected.
3. *Split of the algorithm*, where different parts of the algorithm are executed on different devices. For instance, the solver is executed on one GPU and the preconditioner on another one. In this way the communication between GPUs will be minimized. However this approach requires an individual solution for each algorithm.

Note that the data-parallel approach can be seen as a method splitting the data across multi-GPUs, whereas the split of the algorithm can be seen as a method splitting the tasks across multiple devices. In this paper we are investigating the data-parallel approach and the split of the algorithms and make a comparison between multi-core and multi-GPUs. We leave out the domain decomposition approach because the convergence of the Helmholtz solver is not guaranteed. The data parallel approach is more intuitive and is described in detail in Sect. 4.

3.1 Split of the Algorithm

The split can be unique for every algorithm. The main idea of this approach is to limit communication between GPUs but still be able to compute large problems.

One way to apply this approach to the Bi-CGSTAB preconditioned by shifted Laplace multigrid method is to execute the Bi-CGSTAB on one GPU and the multigrid preconditioner on another one. In this case the communication only between the Krylov solver and preconditioner is required but not for intermediate results.

The second way to apply split of the algorithm to our solver is to execute the Bi-CGSTAB and the finest level of shifted Laplace multigrid across all available GPUs using data parallel approach. The coarser levels of multigrid method are executed on only one GPU due to small memory requirements. Since the LU-decomposition is used to compute an exact solution on the coarsest level, we use the CPU for that.

3.2 Issues

Implementation on multi-GPUs requires careful consideration of possibilities and optimization options. The issues we encountered during our work are listed below:

- Multi-threading implementation, where the life of a thread should be as long as the application. This is crucial for the multi-threading way of implementation on multi-GPU. Note that in case of pushing contexts this is not an issue.
- Because of limited GPU memory size, large problems need multiple GPUs.
- Efficient memory reusage to avoid allocation/deallocation. Due to memory limitations the memory should be reused as much as possible, especially in the multigrid method. In our work we create a pool of vectors on the GPU and reuse them during the whole solution time.
- Limit communications CPU→GPU and GPU→CPU.
- The use of texture memory on Multi-GPU is complicated as each GPU needs its own texture reference.
- Coalescing is difficult since each matrix row has a different number of elements.

4 Numerical Results on Multi-GPU

4.1 Vector- and Sparse Matrix-Vector Operations

Vector operations such as addition, dot product are trivial to implement on multi-GPU. Vectors are split across multiple GPUs, so that each GPU gets a part of the vector. In case of vector addition, the parts of a vector remain on GPU or can be send to a CPU and assembled in a result vector of original size. The speedup for vector addition on 8-GPUs compared to a multi-threaded implementation (12 CPUs) is about 40 times for single and double precision. For the dot product, each GPU sends its own sub-dot product to a CPU, where they will be summed into the final result. The speedup for dot product is about 8 for single precision and 5 for double precision. Note that in order to avoid cache effects on a CPU and to make a fair comparison, the dot product has been taken from two different

Table 1 Matrix-Vector-Multiplication in single (SP) and double (DP) precision

Size	Speedup (SP) 12-cores/1 GPU	Speedup (SP) 12-cores/8 GPUs	Speedup (DP) 12-cores/1 GPU	Speedup (DP) 12-cores/8 GPUs
100,000	54.5	6.81	30.75	5.15
1 Mln	88.5	12.95	30.94	5.97
20 Mln	78.87	12.13	32.63	6.47

Table 2 Speedups for Bi-CGSTAB in single (SP) and double (DP) precision

Size	Speedup (SP) 12-cores/1 GPU	Speedup (SP) 12-cores/8 GPUs	Speedup (DP) 12-cores/1 GPU	Speedup (DP) 12-cores/8 GPUs
100,000	12.72	1.27	9.59	1.43
1 Mln	32.67	7.58	15.84	5.11
15 Mln	45.37	15.23	19.71	8.48

vectors. The speedups for vector addition and dot product on multi-GPU are smaller compared to the single GPU because of the communication between CPU and multiple GPUs.

The matrix is stored in a CRS matrix format (Compressed Row Storage, see e.g. [1]) and is split row-wise. In this case a part of the matrix rows is transferred to each GPU as well as the whole vector. After matrix-vector multiplication parts of the result are transferred to a CPU where they are assembled into the final resulting vector. The timings for matrix-vector multiplication are given in Table 1.

4.2 Bi-CGSTAB and Gauss-Seidel on Multi-GPU

Since the Bi-CGSTAB algorithm is a collection of vector additions, dot products and matrix-vector multiplications described in the previous section, the multi-GPU version of the Bi-CGSTAB is straight forward. In Table 2 the timings of Bi-CGSTAB on many-core CPU, single GPU and multi-GPU are presented. The stopping criterion is 10^{-5}. It is easy to see that the speedup on multi-GPUs is smaller than on a single GPU due to the data transfer between CPU and GPU. Note that for the largest problem in Table 2 it is not possible to compute on a single GPU because there is not enough memory available. However it is possible to compute this problem on multi-GPUs and the computation on multi-GPU is still many times faster than 12-core Westmere CPU.

As mentioned above, the shifted Laplace multigrid preconditioner consists of a coarse grid correction based on the Galerkin method with matrix-dependent

Table 3 Speedups for colored Gauss-Seidel method on different architectures in single precision

Size	12-cores/1 GPU	12-cores/8 GPUs
5 Mln	16.5	5.2
30 Mln	89.1	6.1

prolongation and of a Gauss-Seidel smoother. The implementation of coarse grid correction on multi-GPU is straight forward, since the main ingredient of the coarse grid correction is the matrix-vector multiplication. The coarse grid matrices are constructed on a CPU and then transferred to the GPUs. The matrix-vector multiplication on multi-GPU is described in Sect. 4.1.

The Gauss-Seidel smoother on multi-GPU requires adaptation of the algorithm. We use 8-colored Gauss-Seidel, since the problem (1) is given in three dimensions and computations at each discretization point should be done independently of the neighbours to allow parallelism. For the multi-GPU implementation the rows of the matrix for one color will be split between multi-GPUs. Basically, the colors are computed sequentially, but within a color the data parallelism is applied across the multi-GPUs. The timing comparisons for 8-colored Gauss-Seidel implementation on different architectures are given in Table 3.

5 Numerical Experiments for the Wedge Problem

This model problem represents a layered heterogeneous problem taken from [3]. For $\alpha \in \mathbb{R}$ find $\phi \in \mathbb{C}^{n\times n\times n}$

$$-\Delta\phi(x,y,z) - (1-\alpha i)k(x,y,z)^2\phi(x,y,z) = \delta\left((x-500)(y-500)z\right), \quad (2)$$

$(x,y,z) \in \Omega = [0,0,0] \times [1000,1000,1000]$, with the first order boundary conditions. We assume that $\alpha = 0$. The coefficient $k(x,y,z)$ is given by $k(x,y,z) = 2\pi f l/c(x,y,z)$ where $c(x,y,z)$ is presented in the Fig. 2. The grid size satisfies the condition $\max_x(k(x,y,z))h = 0.625$, where $h = \frac{1}{n-1}$. Table 4 shows timings for Bi-CGSTAB preconditioned by the shifted Laplace multigrid method on the problem (2) with 43 millions unknowns. The single GPU implementation is about 13 times faster than a multi-threaded CPU implementation. The data-parallel approach shows that on multi-GPUs the communication between GPUs and CPUs takes a significant amount of the computational time, leading to smaller speedup than on a single GPU. However, using the split of the algorithm, where Bi-CGSTAB is computed on one GPU and the preconditioner on the another one, increases the speedup to 15.5 times. Figure 3 shows the real part of the solution for 30 Hz.

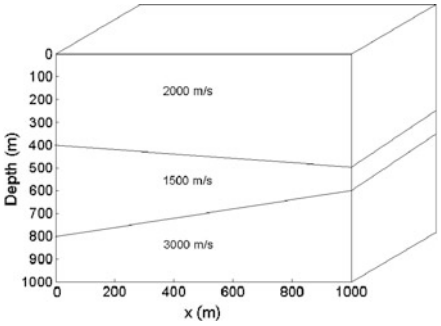


Fig. 2 The velocity profile of the wedge problem

Table 4 Timings for Bi-CGSTAB preconditioned by shifted Laplace multigrid

Size	12-cores/1 GPU Bi-CGSTAB (DP)	12-cores/8 GPUs Preconditioner (SP)	Total	Speedup
12-cores	94 s	690 s	784 s	1
1 GPU	13 s	47 s	60 s	13.1
8 GPUs	83 s	86 s	169 s	4.6
2 GPUs+split	12 s	38 s	50 s	15.5

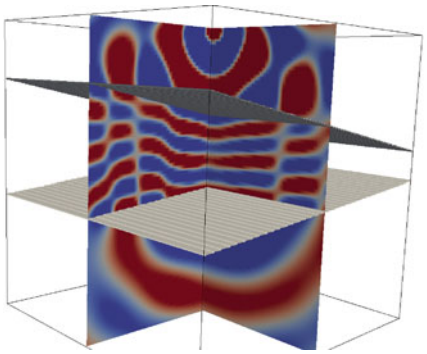


Fig. 3 Real part of the solution, $f = 30$ Hz

6 Conclusions

In this paper we presented a multi-GPU implementation of the Bi-CGSTAB solver preconditioned by a shifted Laplace multigrid method for a three-dimentional Helmholtz equation. To keep the double precision convergence the Bi-CGSTAB method is implemented on GPU in double precision and the preconditioner in

single precision. We have compared the multi-GPU implementation to a single-GPU and a multi-threaded CPU implementation on a realistic problem size. Two multi-GPU approaches have been considered: data parallel approach and a split of the algorithm. For the data parallel approach, we were able to solve larger problems than on one GPU and get a better performance than multi-threaded CPU implementation. However due to the communication between GPUs and a CPU the resulting speedups have been considerably smaller compared to the single-GPU implementation. To minimize the communication but still be able to solve large problems we have introduced split of the algorithm. In this case the speedup on multi-GPUs is similar to the single GPU compared to the multi-core implementation.

The autors thank NVIDIA Corporation for access to the latest many-core-multi-GPU architecture.

References

1. J.J. Dongarra, I.S. Duff, D.C. Sorensen, and H.A. van der Vorst. *Solving Linear Systems on Vector and Shared Memory Computers*. SIAM, Philadelphia (1991).
2. B. Engquist and A. Majda. Absorbing boundary conditions for numerical simulation of waves. *Math. Comput.*, 31:629–651 (1977).
3. Y. A. Erlangga, C. W. Oosterlee, and C. Vuik. A novel multigrid based preconditioner for heterogeneous Helmholtz problems. *SIAM J. Sci. Comput.*, 27:1471–1492 (2006).
4. O. Ernst and M. Gander. Why it is difficult to solve Helmholtz problems with classical iterative methods. In *Durham Symposium 2010* (2010).
5. H. Knibbe, C. W. Oosterlee, and C. Vuik. GPU implementation of a Helmholtz Krylov solver preconditioned by a shifted Laplace multigrid method. *Journal of Computational and Applied Mathematics*, 236:281–293 (2011).
6. www.nvidia.com (2011).
7. E. Zhebel. *A Multigrid Method with Matrix-Dependent Transfer Operators for 3D Diffusion Problems with Jump Coefficients*. PhD thesis, Technical University Bergakademie Freiberg, Germany (2006).

CUDA-Based Parallel Preconditioning for RANS Simulations of Indoor Airflow

S.C. Kramer, C. Pfaffenbach, and G. Lube

Abstract We describe a CUDA-based parallel preconditioning method for non-normal matrices. In particular, we are interested in solving the non-isothermal Reynolds-averaged Navier-Stokes equations. These are at the bottom of indoor air-flow simulations which are necessary for predicting the energy consumption of a building configuration. Within each timestep one has to solve linearized auxiliary problems of Oseen and advection-diffusion-reaction type. Solving the linear algebraic subproblems is accelerated by CUDA by nearly an order of magnitude. Particularly suited is the sparse approximate inverse approach which yields promising results.

1 Introduction

The accurate numerical prediction of indoor-air flows for building configurations of practical relevance [6] is of paramount importance for the energy-efficient design of modern buildings. A major operation in such computations is the solution of large linear algebraic systems of equations arising from the discretized flow problem.

With the advent of multicore processors and many-core, programmable graphics cards numerical linear algebra can be further accelerated using the fine-grained parallelism of CUDA (compute unified device architecture by NVIDIA). Porting iterative solvers to graphics processing units (GPU) is easy, finding a preconditioning strategy matching the data-parallel architecture of nowadays GPUs is not. Particularly suitable are sparse approximate inverses (SpAI) as they only need sparse matrix-vector (SpMV) products for application. To our knowledge

S.C. Kramer (✉) · G. Lube · C. Pfaffenbach
Institut f. Numerische und Angewandte Mathematik, D-37083 Göttingen,
Lotzestrasse 16–18, Germany
e-mail: stkramer@math.uni-goettingen.de; lube@math.uni-goettingen.de;
c.pfaffenbach@math.uni-goettingen.de

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_70,

there has been published only little about CUDA-based parallel preconditioners for the nonsymmetric case and even less for matrices arising from multiphysics applications like non-isothermal air-flow. For factorization-based preconditioners there exist a parallel implementation of block-diagonal ILU [8], ILU [4] and a biorthogonalization-based SpAI [9] which is the work most closely related to ours. Unlike us, [9] only measures the speedup of a CUDA-based implementation of SpAI over an OpenMP-based one for different sparsification strategies. In contrast to that we compare the performance of a serial ILU-implementation with an unfactored SpAI. Sparse approximate inverses can tackle indefinite matrices as well and thus have a broader scope of applicability than ILU. Section 2 summarizes the computational model for low-turbulent indoor airflow and its discretization. In Sect. 3 we give an outline of SpAI and Sect. 4 contains the essential details of its CUDA implementation. Numerical results are presented in Sect. 5 and conclusions in Sect. 6.

2 Bouyancy Driven Fluid Flow

The basis of indoor airflow simulations are the non-dimensional incompressible, non-isothermal Reynolds-Averaged Navier-Stokes (RANS) equations as presented in detail earlier in [5] which also covers a more general set of boundary conditions, shock capturing and the details of the wall function method we employ. Buoyancy forces are modeled by the Boussinesq approximation. In a bounded domain $\Omega \subset \mathbb{R}^3$ with boundary $\partial\Omega$ we seek velocity $\mathbf{u}$, pressure p, and temperature θ solving

$$\begin{aligned} \partial_t \mathbf{u} - \nabla \cdot (2\nu_e \mathbb{S}(\mathbf{u})) + (\mathbf{u} \cdot \nabla)\mathbf{u} + \nabla p &= -\beta\theta\mathbf{g}\,, \\ \nabla \cdot \mathbf{u} &= 0\,, \\ \partial_t \theta + (\mathbf{u} \cdot \nabla)\theta - \nabla \cdot (a_e \nabla\theta) &= c_p^{-1}\dot{q}^V \end{aligned} \tag{1}$$

with the rate of strain tensor $\mathbb{S}(\mathbf{u}) := (\nabla\mathbf{u} + \nabla\mathbf{u}^T)/2$, isobaric volume expansion coefficient β, gravitational acceleration $\mathbf{g}$, volumetric heat source $\dot{q}^V$, and isobaric specific heat capacity c_p. We introduce effective viscosities $\nu_e = \nu + \nu_t$ and $a_e = a + a_t$ with kinematic viscosity ν, turbulent viscosity ν_t, thermal diffusivity $a = \nu/Pr$ and turbulent thermal diffusivity $a_t = \nu_t/Pr_t$ with Prandtl numbers $Pr = 0.7$ and $Pr_t = 0.9$ for air. The non-constant ν_t and a_t reflect turbulent effects and depend on the turbulence model. The sign of $\mathbf{u} \cdot \mathbf{n}$, $\mathbf{n}$ being the outer normal, rules the division of $\partial\Omega$ into wall zones Γ_W, inlet zones Γ_- and outlet zones Γ_+ where we impose

$$\sigma(\mathbf{u}, p)\mathbf{n} \equiv 2\nu_e \mathbb{S}(\mathbf{u}) - p\mathbb{I} = \tau_n \mathbf{n} \ \text{ on } \Gamma_- \cup \Gamma_+, \qquad \mathbf{u} = \mathbf{0} \ \text{ on } \Gamma_W, \tag{2}$$

$$\theta = \theta_{in} \ \text{ on } \Gamma_-, \ \ a_e \nabla\theta \cdot \mathbf{n} = 0 \ \text{ on } \Gamma_+, \quad \theta = \theta_w \ \text{ on } \Gamma_W. \tag{3}$$

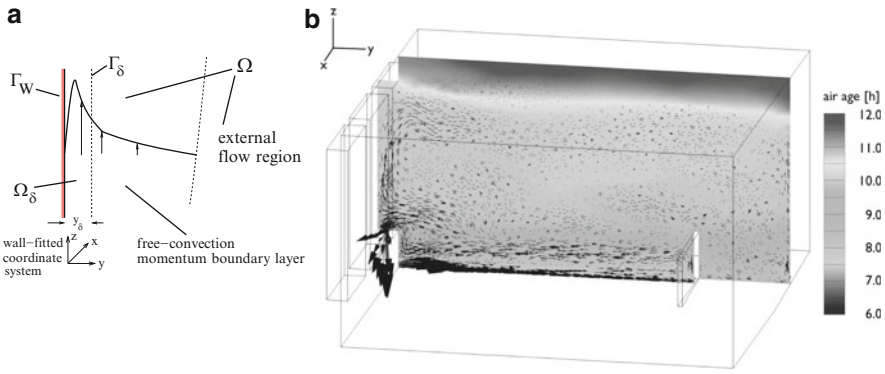


Fig. 1 (**a**) Separate treatment of boundary layer Ω_δ. (**b**) Flow field and distribution of air age

The in- and outflow conditions in Eqs. (2) and (3) are suitable for natural ventilation. Near Γ_W, $\mathbf{u}$ and θ exhibit strong gradients. Figure 1 shows a typical near-wall profile for the streamwise component of $\mathbf{u}$ for the flow along a heated vertical wall. Time discretization is performed with the BDF(1) scheme. This leads to a sequence of coupled nonlinear problems within each time step. The global problem in Ω reads

$$-\nabla\cdot(2\nu_e\mathbb{S}(\mathbf{u}))+(\mathbf{u}\cdot\nabla)\mathbf{u}+\frac{1}{\delta t}\mathbf{u}+\nabla p=-\beta\theta\mathbf{g}+\frac{1}{\delta t}\mathbf{u}_{old},\quad \nabla\cdot\mathbf{u}=0, \tag{4}$$

$$-\nabla\cdot(a_e\nabla\theta)+(\mathbf{u}\cdot\nabla)\theta+\frac{1}{\delta t}\theta=c_p^{-1}\dot{q}^V+\frac{1}{\delta t}\theta_{old} \tag{5}$$

with modified boundary conditions on Γ_W

$$\mathbf{u}\cdot\mathbf{n}=0\,,\quad(\mathbb{I}-\mathbf{n}\otimes\mathbf{n})\sigma(\mathbf{u},p)\mathbf{n}=\tau_t(\mathbf{u},\mathbf{u}^L,\theta^L)\,, \tag{6}$$

$$a_e\nabla\theta\cdot\mathbf{n}=c_p^{-1}\dot{q}(\mathbf{u}^L,\theta^L). \tag{7}$$

Boundary data τ_t, $\dot{q}$ at Γ_W are taken from the solution $(\mathbf{u}^L, p^L, \theta^L)$ of Eq. (4) in the boundary layer $\Omega_\delta := \{x \in \Omega : \mathrm{dist}(x,\Gamma_W) < y_\delta\}$ with a modified wall-function approach, thus avoiding locally fine meshes and expensive anisotropic grid refinement in Ω_δ. For indoor airflow the k-ϵ turbulence model with $\nu_t = c_\mu k^2/\epsilon$, $c_\mu = 0.09$ is a reasonable choice. The turbulent kinetic energy k and dissipation ϵ solve

$$-\nabla\cdot(\nu_k\nabla k)+(\mathbf{u}\cdot\nabla)k+\frac{1}{\delta t}k=P_k+G-\epsilon+\frac{1}{\delta t}k_{old}\,, \tag{8}$$

$$-\nabla\cdot(\nu_\epsilon\nabla\epsilon)+(\mathbf{u}\cdot\nabla)\epsilon+\frac{1}{\delta t}\epsilon+C_2\frac{\epsilon^2}{k}=C_1\frac{\epsilon^2}{k}(P_k+G)+\frac{1}{\delta t}\epsilon_{old} \tag{9}$$

with effective viscosities $\nu_k = \nu + \nu_t/Pr_k, \nu_\epsilon = \nu + \nu_t/Pr_\epsilon$, production and buoyancy terms $P_k = 2\nu_t|\mathbb{S}(\mathbf{u})|^2$, $G = \beta a_t \mathbf{g} \cdot \nabla\theta$ and empirical constants $C_1 = 1.44, C_2 = 1.92, Pr_k = 1.0, Pr_\epsilon = 1.3$. The k-ϵ Eqs. (8) and (9) are solved in $\Omega \backslash \Omega_\delta$ with appropriate boundary conditions for k, ϵ on Γ_δ. For the full discretization we decouple and linearize the model within each time step. Two basic problems are to be solved: (i) An Oseen problem with variable viscosity ν and positive reaction term

$$\begin{aligned} -\nabla \cdot (2\nu\mathbb{S}(\mathbf{u})) + (\mathbf{a} \cdot \nabla)\mathbf{u} + c\mathbf{u} + \nabla p &= \mathbf{f} && \text{in } \Omega\,, \\ \nabla \cdot \mathbf{u} &= 0 && \text{in } \Omega \end{aligned} \tag{10}$$

with $\sigma(\mathbf{u}, p)\mathbf{n} = \tau_n \mathbf{n}$ on $\Gamma_- \cup \Gamma_+$, $(\mathbb{I} - \mathbf{n} \otimes \mathbf{n})\sigma(\mathbf{u}, p)\mathbf{n} = \tau_t$ and $\mathbf{u} \cdot \mathbf{n} = 0$ on Γ_W. (ii) Advection-diffusion-reaction (ADR) problems for θ, k and ϵ with variable viscosity

$$-\nabla \cdot (\nu \nabla u) + (\mathbf{a} \cdot \nabla)u + cu = f \qquad \text{in } \tilde{\Omega} \tag{11}$$

where $\tilde{\Omega} = \Omega$ or $\tilde{\Omega} = \Omega \setminus \Omega_\delta$ with Dirichlet boundary $\tilde{\Gamma}_D$ and von Neumann boundary $\tilde{\Gamma}_N$. The boundary conditions are $u = g$ on $\tilde{\Gamma}_D$ and $\nu \nabla u \cdot \mathbf{n} = h$ on $\tilde{\Gamma}_N$. The testcase considered in Sect. 5 requires an additional equation like (11) for the air age.

For the finite element discretization of (10) and (11) we assume an admissible triangulation of Ω and define discrete subspaces of globally continuous and piecewise linear ansatz and test functions. The standard Galerkin FEM for the Oseen problem (10) with an equal-order ansatz for velocity and pressure does not pass the discrete inf-sup condition and thus must be stabilized. The Galerkin-FEM for the ADR-problem (11) needs stabilization, too. The resulting linear systems are highly non-symmetric and in general of non-normal type. The discretized Oseen problem has a saddle-point structure. Figure 2 shows the dominating eigenvalues of the discretized Oseen problem for the testcase to be considered in Sect. 5. Preconditioning strategies for Eqs. (10) and (11) aim at a better clustering of the spectra.

3 Preconditioning Strategies

Given a linear system $Ax = b$ with $A \in \mathbb{R}^{n \times n}$ and $x, b \in \mathbb{R}^n$ where x is the sought solution, the essence of preconditioning is to find a matrix $M \in \mathbb{R}^{n \times n}$ of which the inverse M^{-1} is easy to compute and yet approximates the inverse A^{-1} well. The iterative solution is obtained either from the right- or left-preconditioned system

$$AM^{-1}y = b \qquad x = M^{-1}y\,, \tag{12}$$

$$M^{-1}Ax = M^{-1}b\,. \tag{13}$$

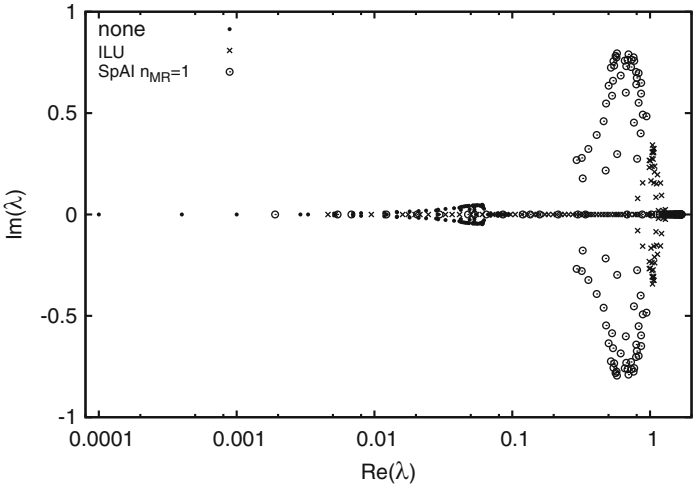


Fig. 2 The dominating 128 eigenvalues λ of the unpreconditioned Oseen matrix ($\cdot$), right-preconditioned by ILU ($\times$) and by SpAI ($\circ$). The x-axis is in logarithmic scale to highlight the order of magnitude of the real parts. Eigenvalues were computed from the Hessenberg matrix obtained from the Arnoldi process using MATLAB's `eig()` function

For SpAI inverting M is meant literally and applying M^{-1} means a single SpMV product. To compute a SpAI $M^{-1} \approx A^{-1}$ we solve the n independent minimization problems for the columns m_j of M^{-1} where e_j is the jth column unit vector

$$M^{-1} := \underset{S \in \mathbb{R}^{n \times n}}{\arg\min} \|I - AS\|_F\,, \qquad m_j := \underset{s \in \mathbb{R}^n}{\arg\min} \|e_j - As\|_2\,. \tag{14}$$

Equations (14) are solved iteratively using the well-known minimal residual (MinRes) algorithm with initial guess $M_0^{-1} = A^T$. We compute all m_j in parallel by

$$r_j = e_j - Am_j \tag{15}$$

$$\alpha_j = \frac{(r_j, Ar_j)}{(Ar_j, Ar_j)} \tag{16}$$

$$m_j = m_j + \alpha_j r_j \tag{17}$$

until convergence. The SpMV products Ar_j are computed only once per step. The residuals r_j can be computed simultaneously by the corresponding sparse-matrix-sparse-matrix products. Similarly, all α_j can be computed in parallel easily, as there are no dependencies among the scalar products nor among the different α_j. Even though A and M^{-1} are sparse, the matrix $R \equiv (r_0, \ldots, r_{n-1})$ might be not. The same holds for $P \equiv (p_0, \ldots, p_{n-1}) = AR$. To avoid a huge fill-in we fix the sparsity patterns of both R and P beforehand to be the one of A^T.

4 Implementation Issues

A general feature of finite-element matrices on arbitrary tetrahedral meshes is a lack of structure. We compensate this by ordering the solution components per grid point so that in our vector-valued subproblems (Oseen and k-ϵ) some local structure is induced and matrix elements can be stored as small dense matrices. Then, multiplying a matrix element with a vector element is itself a product of a dense matrix and vector which due its regular memory access pattern matches the GPU architecture well. Due to this special structure we implemented our own SpMV product which we have based on [2] and reach roughly 50–70% of the peak memory bandwidth.

The essential operation for assembling a SpAI is the inner product of two sparse vectors whereas its application only requires the standard SpMV product. Due to the ordering of the solution components for vector-valued problems multiplying two vector elements amounts to multiplying small dense matrices. Especially in the 4×4, i.e. Oseen, case this allows for an efficient hardware utilization by assigning each inner product to a half-warp. Then, global memory accesses in loading source and destination elements fully coalesce as they take place as multiples of 128 byte which is an exact match for the size of the cache lines [1]. Our inner product resembles the one given in [9] except that we have to multiply and add small dense matrices.

5 Numerical Results

Our test matrices stem from a case study of indoor air flow in a room subject to energy-focussed building refurbishment. The goal was to measure the impact on air exchange by decentralized air-conditioning attached to the windows. The room was discretized by a tetrahedral mesh with 80,621 nodes resulting in 644,968 degrees of freedom (DoFs) in total. For details see Table 1. A snapshot of the flow field and distribution of air age is displayed in Fig. 1. We solve Eqs. (10) and (11) by GMRES [7] and QMRCGSTAB [3]. Especially the latter has turned out to be the best choice for solving a broad range of problems. On the CPU we use an ILU(0) preconditioner. For judging performance we compared the spectra of the preconditioned linear systems and convergence history. For good preconditioning eigenvalues should cluster in the vicinity of $(1, 0)$, cf. Fig. 2. The convergence histories for different numbers n_{MR} of MinRes steps are summarized in Fig. 3. Total runtimes and times per iteration are given in Table 1.

Our tests were done on a Dell T7500 workstation equipped with two Intel Xeon X5650, 96 GB RAM, and two NVIDIA Tesla C2070, running under Ubuntu 10.04 and CUDA 4.0. As integrated development environment we use QtCreator. For performance tests programs were compiled in its predefined release mode. Our tests show that already $n_{MR} = 1$ suffices to obtain a preconditioner with reasonable performance, cf. Fig. 3. SpAI takes more steps to converge than ILU but this is

Table 1 Typical SpAI speedups and parallelization strategy. To optimize usage of CPUs and GPU move the Oseen problem to the GPU and solve the other problems on the CPU

Case	nnz	Dim	DoFs	Iterations to convergence		Time per iteration/s	Walltime/s
				SpAI	1280	0.014	18
Oseen	18,275,344	4	322,484	ILU	180	0.6	108
				SpAI	60	0.01	0.6
Fourier	1,142,209	1	80,621	ILU	8	0.05	0.4
				SpAI	240	0.01	2.4
k-ϵ	4,568,836	2	161,242	ILU	5	0.20	1
				SpAI	64	0.01	0.6
Air-age	1,142,209	1	80,621	ILU	45	0.05	2.5

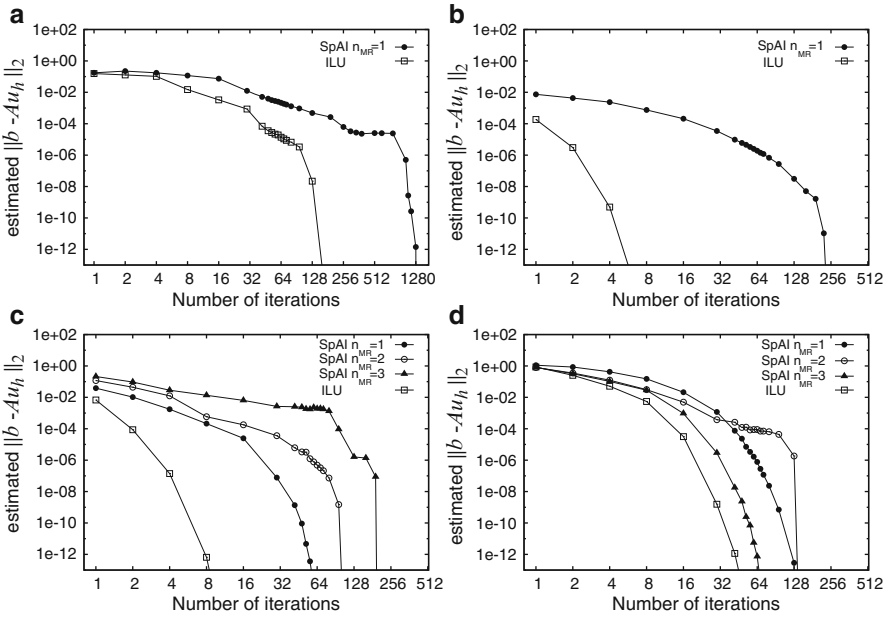


Fig. 3 SpAI results for QMRCGSTAB [3]. In contrast to GMRES convergence is faster by ca. 1.5. (**a**) Oseen. (**b**) k-ϵ. (**c**) Fourier. (**d**) Air age

more than compensated by the efficient implementation of the SpMV which is needed to apply the preconditioner. Table 1 shows that for the Oseen subproblem, which is the most expensive one, we measured an average speed up of 43 for the individual iteration step but have to pay with a seven times higher iteration count until convergence. Thus, our CUDA-based SpAI is six times faster than the ILU preconditioner used so far. As the increased iteration count is due to the change of the preconditioner from ILU to SpAI the same behavior should occur if the SpAI was parallelized on the CPU using OpenMP for instance. For that case we would

expect a speedup of the individual iteration step which would only compensate the increased amount of iterations due to the lower memory bandwidth of current CPUs compared to GPUs so that there would be no net gain. For instance, a Westmere Xeon has roughly one-fifth of the bandwith of a Tesla card. Hence, to get a speed up due to massive parallelization a high-bandwidth architecture like CUDA is mandatory.

Not all of the subproblems profit from switching to SpAI-based preconditioning, especially the turbulence model shows an unfavorable convergence behavior compared to preconditioning by ILU. Thus, we suggest a hybrid strategy where one host thread manages the Oseen problem and its solution on the GPU and where a second host thread solves the turbulence model, Fourier law and air age on the CPU.

With respect to the number of iterations SpAI is not as effective as ILU. Yet, ILU is outperformed because the individual SpAI-preconditioned Krylov iteration step is almost two orders of magnitudes faster, especially for the Oseen problem. Eq. (12) shows that further improvements must tackle the problem of finding a cheap inverse of AM^{-1}. Due to our severe dropping strategy, cf. end of Sect. 3, it does not make sense to replace A by AM^{-1} in Eq. (14). A simple and cheap improvement would be to use a block-Jacobi preconditioner based on the diagonal elements $D_i := (AM^{-1})_{ii}$ and to precondition Eq. (12) with the matrix $J^{-1} := (D_i^{-1})_{i=0}^{N}$ from the left. The structure of J^{-1} should simplify coalesced memory accesses in the SpMV thus exploiting the theoretical memory bandwidth.

6 Conclusion and Outlook

We considered CUDA-based parallel preconditioning for non-normal matrices as they arise from fluid mechanics problems in engineering indoor airflow. We used the sparse approximate inverse approach since it provides some built-in parallelism already by design. A low-turbulent indoor airflow case study served as testbed.

Our results show that sparse approximate inverses can outperform our CPU-based ILU preconditioner by almost an order of magnitude despite a higher iteration count. A comparison of spectra shows that a sparse approximate inverse restricted to the sparsity pattern of the transpose matrix leads to a similar clustering of eigenvalues as ILU does. Therefore, we can accelerate the algebraic part of turbulent indoor air flow simulations by almost an order of magnitude by moving the expensive Oseen problem to the GPU and to use SpAI as preconditioner while solving all other problems in a separate thread on the CPU using ILU as preconditioner.

Acknowledgements We thank NVIDIA corporation for financial and material support and R. Gritzki and M. Rösler for providing us with the problem and preparing the test matrices.

References

1. Fermi compute architecture whitepaper. Technical report, Nvidia Corporation, 2009.
2. N. Bell and M. Garland. Efficient sparse matrix-vector multiplication on CUDA. NVIDIA Technical Report NVR-2008-004, NVIDIA Corporation, 2008.
3. T. F. Chan, E. Gallopoulos, V. Simoncini, T. Szeto, and C. H. Tong. A quasi-minimal residual variant of the bi-cgstab algorithm for nonsymmetric systems. *SIAM J. Sci. Comput.*, 15:338–347, 1994.
4. V. Heuveline, D. Lukarski, and J.-P. Weiss. Fine-grained Parallel ILU Preconditioners with Fill-ins for Multi-core CPUs and GPUs. In *International Conference On Preconditioning Techniques For Scientific And Industrial Applications 2011*, accepted.
5. T. Knopp, G. Lube, R. Gritzki, and M. Rösler. A near-wall strategy for buoyancy-affected turbulent flows using stabilized fem with applications to indoor air flow simulation. *Computer Methods in Applied Mechanics and Engineering*, 194(36–38):3797 – 3816, 2005.
6. G. Lube, T. Knopp, G. Rapin, R. Gritzki, and M. Rösler. Stabilized finite element methods to predict ventilation efficiency and thermal comfort in buildings. *International Journal for Numerical Methods in Fluids*, 57(9):1269–1290, 2008.
7. Y. Saad. *Iterative methods for sparse linear systems*. Second edition, 2003.
8. M. Wang, H. Klie, M. Parashar, and H. Sudan. Solving sparse linear systems on nvidia tesla gpus. In *Proceedings of the 9th International Conference on Computational Science: Part I*, ICCS '09, pages 864–873, Berlin, Heidelberg, 2009. Springer-Verlag.
9. S. Xu, W. Xue, K. Wang, and H. X. Lin. Generating approximate inverse preconditioners for sparse matrices using cuda and gpgpu. *Journal of Algorithms & Computational Technology*, 5(3):475 – 500, 2011.

Shallow Water Simulation on GPUs for Sparse Domains

M.L. Sætra

Abstract Efficient stencil operations are essential in explicit schemes for evolutionary PDEs. In particular, for conservation and balance laws, the solution will in many cases have non-constant values only in a portion of the grid. We present novel methods that through simple observation of the stencil and the distribution of conserved quantities, reduce both the memory footprint and the computational burden by only computing in cells in which the solution changes. To this end, we utilize sparse updating of grid cells, in which data values are not stored before they actually contribute in the simulation. This is motivated by the need to perform simulations over very large domains to model real-world dam breaks and various flooding scenarios. The methods are applied to a high-resolution shallow water simulator, but are also applicable to other stencil-based explicit solvers.

1 Introduction

The graphics processing unit (GPU) has over the last decade been introduced into high-performance computing (HPC) because of its unprecedented floating-point performance for highly parallel code [1, 3]. Today, three of the top five supercomputers in the world utilize GPUs to accelerate computations [6]. Over the past decade, the focus has shifted from simply getting an algorithm to run correctly on the GPU to fully utilizing the GPU hardware and to algorithmic development in general. This shift came as a natural consequence of the maturation of GPU computing as a research field, the increasing adoption of GPU hardware within both research- and HPC-communities, and the availability of tools and libraries. CUDA [7] is of special significance in this regard, which our work is based on.

M.L. Sætra (✉)
Centre of Mathematics for Applications, University of Oslo, P.O. Box 1053 Blindern, NO-0316 Oslo, Norway
e-mail: m.l.satra@cma.uio.no

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_71,

Numerous papers has been published on GPU implementations of explicit schemes for conservation and balance laws, in particular for the shallow water equations, see Brodtkorb et al. [2] and references therein. Most of these approaches perform computations on all cells and do not exploit that large parts of the domain typically do not require any computations for (long) periods of the simulations. Brodtkorb et al. [2] present, to the best of our knowledge, the first attempt to utilize sparsity of cell updates, and the work presented here is a continuation of this work. We present two novel algorithms; both have increased performance and one conserves memory as well. The algorithms are also applicable to other explicit schemes.

We start in Sect. 2 by introducing the shallow water equations and our simulator. Section 3 details our algorithms for utilizing sparsity of cell updates. Results are given in Sect. 4 before we conclude with a short summery.

2 Shallow Water Simulation

The shallow water equations are derived by depth-averaging the Navier-Stokes equations, and are used to model flows where the horizontal scale is much larger than the vertical scale. The equations describe gravity-induced motion, and can capture many naturally occurring phenomena, e.g., tsunamis, inundations, and flash floods. In two dimensions, the shallow water equations with bottom shear stress can be expressed as

$$\begin{bmatrix} h \\ hu \\ hv \end{bmatrix}_t + \begin{bmatrix} hu \\ hu^2 + \frac{1}{2}gh^2 \\ huv \end{bmatrix}_x + \begin{bmatrix} hv \\ huv \\ hv^2 + \frac{1}{2}gh^2 \end{bmatrix}_y = \begin{bmatrix} 0 \\ -ghB_x \\ -ghB_y \end{bmatrix} + \begin{bmatrix} 0 \\ -gu\sqrt{u^2+v^2}/C_z^2 \\ -gv\sqrt{u^2+v^2}/C_z^2 \end{bmatrix}, \tag{1}$$

in which h is the water depth and hu and hv are the discharges along the abscissa and ordinate, respectively. Furthermore, g is the gravitational constant, B is the bottom topography measured from a given datum, and C_z is the Chézy friction coefficient.

Based on this model, we have developed a simulator that has been successfully verified against analytical solutions and validated against real-world dam breaks [2]. We use a second-order, semi-discrete, finite-volume scheme developed by Kurganov and Petrova [5] for the spatial discretization, and solve the resulting ODEs by simple first-order Euler integration or by a second-order, total-variation diminishing Runge-Kutta method. The scheme supports dry zones and is particularly well suited for implementation on the GPU.

The numerical scheme is implemented using four CUDA *kernels* that perform flux calculations, compute the size of the timestep, do time integration, and apply boundary conditions, respectively. This partitioning of the application was found by minimizing the number of kernels while still obeying the data dependencies of the

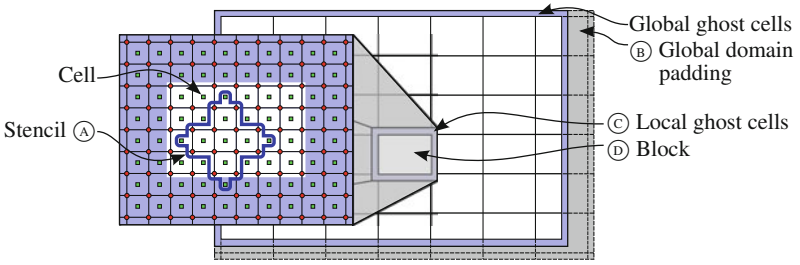


Fig. 1 Domain decomposition and stencil. The global domain is padded Ⓑ to fit an integer number of blocks. Each block Ⓓ has local ghost cells Ⓒ that overlap with other blocks to satisfy the data dependencies dictated by the stencil Ⓐ

numerical scheme. The flux-calculation kernel computes fluxes and source terms, and determines the largest eigenvalues of a portion of the grid. This is the most compute intensive kernel, comprising over 80 % of the total runtime. The timestep kernel performs a standard parallel reduction of the local maximum eigenvalues and then uses the global maximum eigenvalue to determine the size of the timestep to be used by the time-integration kernel to advance the solution in time. Last, the boundary-condition kernel can be executed to update the boundary cells according to given boundary conditions.

Structured grids fit the execution model of the GPU perfectly; one CUDA thread treats one cell in the structured grid (see Fig. 1). A CUDA kernel is executed using thousands or millions of threads organized in a 2D *grid*. This grid is further partitioned into 2D *blocks*, each containing a certain number of threads chosen at runtime, in our case 16×12 threads. When solving hyperbolic (and parabolic) PDEs, explicit methods can be applied because the propagation of the solution is finite and bounded by a CFL-condition. In conjunction with initial perturbations that do not cover the whole domain, this means that it is not necessary to compute all the cells in the domain every timestep of the simulation. This is what enables the algorithms described in this paper, which are motivated by the need for increased performance and domain size.

3 Sparse Cell Updates

To exploit sparsity of cell updates, we need to know which cells contributed to the solution in the previous timestep. The time-integration kernel marks all cells in which the conserved quantity is larger than some predefined ϵ as *wet*. A parallel reduction is then performed per block, and if the block has one or more wet cells, the whole block is marked as wet in a 2D boolean map of all the blocks composing the domain. Since we do not know a priori whether the solution will propagate over a block boundary during the next timestep, we have to include all neighbors of contributing blocks as well (each block has overlapping local ghost cells shared

with its four neighbors in the x- and y-directions). This is done at the granularity of blocks since cancelling out single cells would not fit the execution model of the GPU. Likewise, cells in which the fluxes are perfectly balanced with the source terms are also excluded, meaning that a lake-at-rest will not be updated before some disturbances are introduced.

Brodtkorb et al. [2] proposed an *early exit* of blocks that did not contribute to the solution in the previous timestep. Early exit is implemented in the flux and the time-integration kernels by marking wet blocks in the time-integration kernel, and then using the 2D boolean map of wet blocks as input to the flux kernel in the next timestep. Before doing any computations, the flux kernel checks whether the current block and all its neighbors were marked as wet in the previous timestep. If neither the block nor any of its neighbors were marked as wet, all threads in the block exit before performing any computations. There are some penalties to pay for using early exit: (i) the extra reading and writing to the 2D boolean map and (ii) the shared memory, parallel reduction in the time-integration kernel, both adding additional latency to each timestep. The size of the map used for storing each block's state equals (domain width/block width) x (domain height/block height) and is typically only a fraction of the size of the total domain. Hence the added memory usage is negligible.

While the early exit technique leads to a significant reduction of computational costs for many real-world cases, it does not reduce memory consumption. To this end, there are three points we can improve upon: Since one CUDA thread is launched per cell in the full grid, we are launching unnecessary threads every timestep for all blocks that will exit early, since they do not contribute to the current timestep, and while thread switching is inexpensive on the GPU, there is a significant latency connected to launching blocks. At the same time we are wasting memory bandwidth since all these blocks must read from GPU global memory before performing the early exit test. Last, we are wasting memory as the full domain is stored in memory. A natural next step would be to only launch the number of blocks that are actually necessary in each timestep; the wet blocks and their neighbors. Diagonal neighbors are not necessary to include because such connections are not considered in the stencil of our chosen numerical scheme. We have implemented two versions of this sparse update algorithm: *sparse compute* and *sparse memory*.

Sparse compute. In this version, the flux and time-integration kernels launch just the necessary number of blocks needed to correctly capture the next timestep by using a one-dimensional look-up map that is updated after every timestep. The look-up map is needed because we now decouple the logical domain and numerical grid from the CUDA grid. The underlying data structures are, however, not changed in any way. Wet blocks are marked in a 2D boolean map by the time-integration kernel, as in early exit. This 2D map of wet blocks is now used as input to a new kernel, the *grow* kernel, that computes which blocks need to be included in the next timestep (the wet blocks and all neighbors of wet blocks) and writes the indices of these blocks to the look-up map. We call this set of blocks for *active* blocks

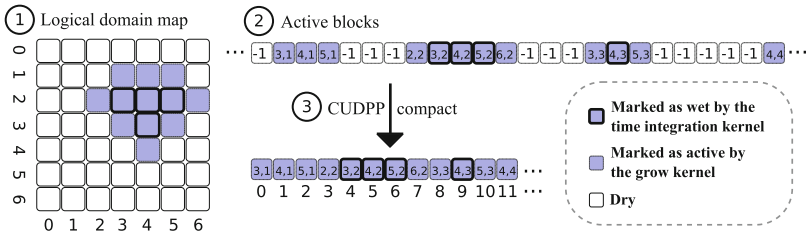


Fig. 2 The time-integration kernel marks the position of all wet blocks in the boolean map ①, and the grow kernel adds all wet blocks, with neighbors, to ②. Both mappings are completely updated every timestep and the 2D coordinates in ②-③ are encoded as single integers

(see ①-② in Fig. 2). By resetting the 2D map of active blocks (to dry) before each timestep, active blocks may become inactive again, although this will rarely happen in practice for the numerical scheme we have chosen. Two copies of the look-up map are needed since the grow kernel outputs one value for every block in the full CUDA grid regardless: the linear index for active blocks and -1 for all other blocks. This first array is then compacted using CUDPP [4], and we get a second array in which the first contiguous elements are the indices of all active blocks (see ③ in Fig. 2). CUDPP also outputs the number of elements in this compacted array, which is the number of active blocks, and thus the number of blocks we need to launch in the next timestep for the flux and time-integration kernels. In the current implementation we launch a one-dimensional CUDA grid as wide as the number of active blocks. This limits the number of active blocks to 65,535 because of constraints in CUDA, and if more blocks are needed, an extension to a 2D CUDA grid would be necessary. The compacted look-up map from the previous timestep is used to find the corresponding data values in memory for all active blocks and contains the linearized 2D index of each active block. Using simple formulas we calculate the 2D coordinates of each active block before loading data in the flux and time-integration kernels.

Sparse memory. This version conserves memory at the cost of some extra complexity in the algorithm. This is achieved by decoupling the logical domain and numerical grid from the underlying data structures. While the logical domain remains the same, the conserved quantities are now stored in a *block-linear* fashion in which the height of the allocated memory is the same as the height of a block, and the width equals the block width times the number of blocks. If the full domain does not fit in the GPU's global memory, it is possible to allocate all available memory and stop the simulation when all allocated memory is in use. This enables simulation of cases that would otherwise not fit in GPU memory. We will now need an additional mapping in order to locate a block's neighbors in the logical full domain. Introducing two more one-dimensional arrays compared with the sparse compute algorithm makes this possible. In this algorithm, wet blocks are not ejected should they become dry again. The complete bathymetry is still loaded at startup, as it only constitutes less than 10 % of the data values stored per cell. By loading the bathymetry on-demand from the CPU (after the grow kernel has

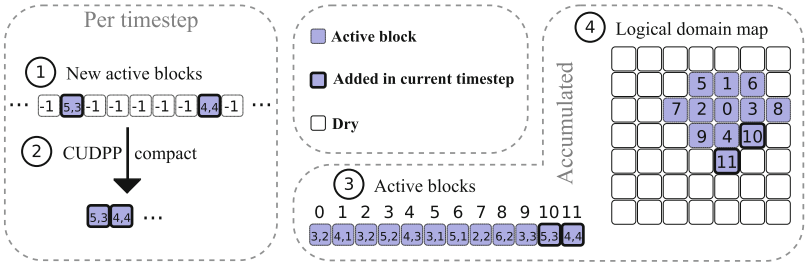


Fig. 3 This figure illustrates an example where two new active blocks are identified by the grow kernel. ① is the output from the grow kernel, and the 2D coordinates in ①-③ are encoded as single integers. The two accumulated look-up maps are updated after each timestep only if new active blocks are found in the current timestep

completed) additional memory could be freed up, but this would be at the expense of performance as the data would have to travel through the PCI Express bus.

The time-integration kernel marks wet blocks by the same criteria as before, but now stores the linearized 2D index of each wet block in a 1D array. Next, the grow kernel adds all new active blocks (new wet blocks and their neighbors) needed in the next timestep by using the array of wet blocks produced by the time-integration kernel, and the 2D map of all active blocks accumulated from all previous timesteps. Blocks that are not already active, and either are wet themselves or have at least one wet neighboring block, are added to an array of new active blocks. This array is then compacted using CUDPP (see ①-② in Fig. 3). If no new active blocks are added, the current timestep is complete. When new active blocks are added, two mappings need to be updated (see ③-④ in Fig. 3): First, a dedicated kernel updates the 2D map of accumulated active blocks. This kernel adds the block-linear memory offsets of the new active blocks in the correct position in the 2D map of all active blocks, relative to the logical domain. The block-linear memory offsets are easily found by iterating from the number of currently active blocks to the number of currently active blocks plus new active blocks. Second, the compacted array of new active blocks that were added in the current timestep is appended to an accumulated map of all active blocks, i.e., the first index in this map contains the linearized 2D coordinates of the first data block stored in memory, and so on.

Prior to a new timestep, the CUDA grid sizes for the flux, time-integration, and timestep-size kernels are adjusted, so the newly added blocks will be included. In the flux kernel, we now need to find each block's neighbors to include the local ghost cells (see Fig. 1). This is done by first finding the block's coordinates in the 2D map, by looking up in the one-dimensional map of all active blocks using CUDA's *blockIdx.x* variable as index, and then use the 2D map to get the block-linear memory offsets of the block's four neighbors (see Fig. 4). Two rows or columns are then read from each neighbor, depending on its relative position. The

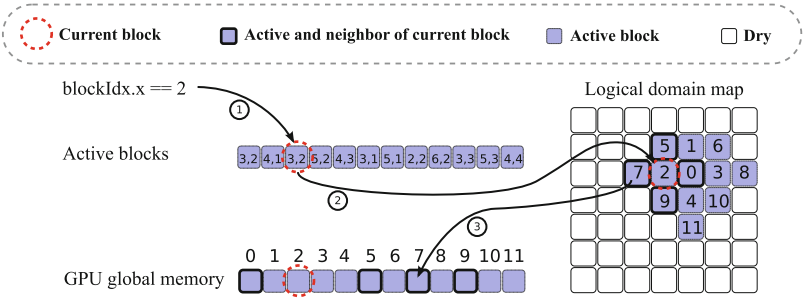


Fig. 4 Illustration of how block with blockIdx.x == 2 finds its west neighbor in the flux kernel: ① the CUDA block index is used to look up in a linear index map of all active blocks to ② find the current block's position in the 2D logical domain. The 2D domain map is then used to find the memory offset of the block's neighbor ③ in block-linear memory. Each block in GPU global memory consists of 16 × 12 cells

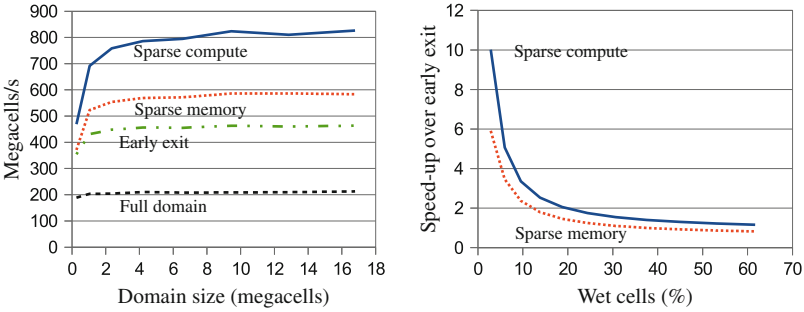


Fig. 5 (*Left*) Comparison of the different algorithms implemented, using second-order Runge-Kutta time integration. (*Right*) A single run using a 4,096×4,096 grid and Runge-Kutta time integration, showing speedup over early exit

time-integration kernel has no inter-block dependencies and can thus simply load the number of currently active blocks from memory, starting with zero offset.

4 Results

An idealized circular dam break is used as a benchmark to demonstrate the performance of the described algorithms. The domain size is 100 m by 100 m, and the initial conditions are a circular dam in the center of the domain with a radius of 6.5 m and a water height of 10 m, whereas the rest of the domain is dry. At timestep zero the dam is removed. The simulation is stopped at 2 s, at which point 53 % of the cells in the full domain are wet.

The left part of Fig. 5 compares the performance between the different algorithms implemented in our simulator as a function of grid resolution. All the algorithms perform better as the number of grid cells is increased and the GPU is saturated with enough threads to efficiently hide latencies and overheads. The performance is close to constant after this happens. As expected, the sparse compute algorithm performs better than the sparse memory algorithm, which requires more bookkeeping. The right part of Fig. 5 shows the performance as the simulation progresses in time and water covers an increasing part of the domain. The sparse memory algorithm becomes slower than early exit at approximately 35 % wet cells. If we consider a real-world example, such as the Malpasset dam break case [2], there are still only 30 % wet cells in the domain after 1 h into the simulation.

From these results we can conclude that sparse compute yields the highest performance increase compared to computing the full domain. On the other hand, if conserving memory is important, then sparse memory should be used, and this algorithm will also give a performance increase over the simple early exit strategy.

5 Summary

We have presented two novel algorithms for sparse cell updates that are applicable to explicit PDE solvers. The efficiency of the algorithms is demonstrated to be excellent for domains which are largely dry, as is the case in many real-world scenarios, such as dam breaks and inundations near riverbanks and coastal regions.

Acknowledgements Part of this work is supported by the Research Council of Norway's project number 180023 (Parallel3D) and the Norwegian Meteorological Institute.

References

1. A. R. Brodtkorb, C. Dyken, T. R. Hagen, J. M. Hjelmervik, and O. Storaasli. State-of-the-art in heterogeneous computing. *Scientific Programming*, 18(1):1–33, May 2010.
2. A. R. Brodtkorb, M. L. Sætra, and M. Altinakar. Efficient shallow water simulations on GPUs: Implementation, visualization, verification, and validation. *Computers & Fluids*, 55(0):1–12, 2012.
3. M. Harris and D. Göddeke. General-purpose computation on graphics hardware. http://gpgpu.org.
4. M. Harris, S. Sengupta, and J. D. Owens. *GPU Gems 3*, chapter 39. Addison-Wesley Professional, first edition, 2007.
5. A. Kurganov and G. Petrova. A second-order well-balanced positivity preserving central-upwind scheme for the Saint-Venant system. *Communications in Mathematical Sciences*, 5:133–160, 2007.
6. H. Meuer, E. Strohmaier, J. Dongarra, and H. Simon. Top 500 supercomputer sites. http://www.top500.org/, December 2011.
7. NVIDIA. NVIDIA CUDA C programming guide 4.1, 2011.

Parallel Implementation of Multilevel BDDC

J. Šístek, J. Mandel, B. Sousedík, and P. Burda

Abstract In application of the Balancing Domain Decomposition by Constraints (BDDC) to a case with many substructures, solving the coarse problem exactly becomes the bottleneck which spoils scalability of the solver. However, it is straightforward for BDDC to substitute the exact solution of the coarse problem by another step of BDDC method with subdomains playing the role of elements. In this way, the algorithm of three-level BDDC method is obtained. If this approach is applied recursively, multilevel BDDC method is derived. We present a detailed description of a recently developed parallel implementation of this algorithm. The implementation is applied to an engineering problem of linear elasticity and a benchmark problem of Stokes flow in a cavity. Results by the multilevel approach are compared to those by the standard (two-level) BDDC method.

J. Šístek (✉)
Institute of Mathematics, Academy of Sciences of the Czech Republic, Žitná 25, Praha 1, CZ-115 67, Czech Republic
e-mail: sistek@math.cas.cz

J. Mandel
Department of Mathematical and Statistical Sciences, University of Colorado Denver, Campus Box 170, Denver, CO 80217-3364, USA
e-mail: jan.mandel@ucdenver.edu

B. Sousedík
Department of Aerospace and Mechanical Engineering, University of Southern California, Olin Hall 430, Los Angeles, CA 90089-2531, USA
e-mail: sousedik@usc.edu

P. Burda
Department of Mathematics, Faculty of Mechanical Engineering, Czech Technical University, Karlovo náměstí 13, Praha 2, CZ-121 35, Czech Republic
e-mail: pavel.burda@fs.cvut.cz

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_72,

1 Introduction

The Balancing Domain Decomposition by Constraints (BDDC) method introduced in [2] is one of the most advanced methods of iterative substructuring for the solution of large systems of linear algebraic equations arising from discretization of boundary value problems. However, in the case of many substructures, solving the coarse problem exactly becomes the limiting factor for scalability of the otherwise perfectly parallel algorithm. This has been observed also for the FETI-DP method (e.g. in [3]), which is closely related to BDDC. For this reason, recent research in the area is directed towards inexact solutions of the coarse problem. For example, algebraic multigrid is used in [3] to obtain an approximate coarse correction within the FETI-DP method, and excellent scalability is achieved.

We follow a different approach in this contribution. As was mentioned already in [2], it is quite straightforward for BDDC to substitute the exact solution of the coarse problem by another step of the BDDC method with subdomains playing the role of elements. In this way, the algorithm of three-level BDDC method is obtained (studied in [8]). If this step is repeated recursively, one arrives at the *multilevel BDDC method* (introduced in [4] without a parallel implementation). Unlike for most other domain decomposition methods, such extension is natural for BDDC, since the coarse problem has the same structure as the original problem. Although the mathematical theory in [4] suggests worsening of the efficiency of the multilevel BDDC preconditioner with each additional level, the resulting algorithm may outperform the standard method with respect to computational time due to better scalability. This fact makes the algorithm a good candidate for using on future massively parallel systems.

In this paper, we present a recently developed parallel implementation of multilevel BDDC method. It is applied to an engineering problem of linear elasticity and a benchmark problem of Stokes flow in a cavity. The results suggest which drawbacks of the two-level implementation might be overcome by the extension to more levels. Our solver library has been released as an open-source package.

2 BDDC Preconditioner with Two and More Levels

The starting point for BDDC is the *reduced interface problem* $\widehat{\mathbf{S}}\,\widehat{\mathbf{u}} = \widehat{\mathbf{g}}$, where $\widehat{\mathbf{S}}$ is the *Schur complement* with respect to *interface*, i.e. unknowns shared by more than one subdomain, $\widehat{\mathbf{u}}$ is the part of vector of coefficients of finite element basis functions at the interface, and $\widehat{\mathbf{g}}$ is sometimes called *condensed right hand side*. This problem is solved by a Krylov subspace method in the framework of *iterative substructuring*. Within these methods, application of $\widehat{\mathbf{S}}$ to a vector is realized by parallel solution of independent *discrete Dirichlet problems*. In this way, the costly explicit construction of the Schur complement is avoided. However, since it is not the main concern of this contribution, the reader is referred to paper [5], or monograph [7] for details of iterative substructuring.

In what follows, we turn our attention towards the second key part of Krylov subspace methods – the *preconditioner*, which is realized by one step of the BDDC method. Let us begin with description of the standard (two-level) version of BDDC. Let $\mathbf{K}_i$ be the local subdomain matrix, obtained by the sub-assembling of element matrices of elements contained in i-th subdomain. We introduce the *coarse space basis functions* on each subdomain represented by columns of matrix Ψ_i, which is the solution to the saddle point problem with multiple right hand sides

$$\begin{bmatrix} \mathbf{K}_i & \mathbf{C}_i^T \\ \mathbf{C}_i & \mathbf{0} \end{bmatrix} \begin{bmatrix} \Psi_i \\ \Lambda_i \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix}. \tag{1}$$

Matrix $\mathbf{C}_i$ represents constraints on functions Ψ_i, one row per each. These constraints enforce continuity of approximate solution at *corners* and/or continuity of more general quantities, such as averages over shared subsets of interface (*edges* or *faces*) between adjacent subdomains. The *local coarse matrix* $\mathbf{K}_{Ci} = \Psi_i^T \mathbf{K}_i \Psi_i = -\Lambda_i$ is constructed for each subdomain. The *global coarse matrix* $\mathbf{K}_C$ is obtained by the assembly procedure from local coarse matrices. This can be formally written as $\mathbf{K}_C = \sum_{i=1}^N \mathbf{R}_{Ci}^T \mathbf{K}_{Ci} \mathbf{R}_{Ci}$, where $\mathbf{R}_{Ci}$ realize the restriction of global coarse degrees of freedom to local coarse degrees of freedom of i-th subdomain.

Suppose $\widehat{\mathbf{r}} = \widehat{\mathbf{g}} - \widehat{\mathbf{S}}\,\widehat{\mathbf{u}}$ is a residual within the Krylov subspace method. The residual assigned to i-th subdomain is computed as $\mathbf{r}_i = \mathbf{E}_i^T \widehat{\mathbf{r}}$, where matrices of weights $\mathbf{E}_i^T$ distribute $\widehat{\mathbf{r}}$ to subdomains. The subdomain correction is now defined as the solution to the system

$$\begin{bmatrix} \mathbf{K}_i & \mathbf{C}_i^T \\ \mathbf{C}_i & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{z}_i \\ \boldsymbol{\lambda}_i \end{bmatrix} = \begin{bmatrix} \mathbf{r}_i \\ \mathbf{0} \end{bmatrix}. \tag{2}$$

The residual for the coarse problem is constructed using the coarse basis functions subdomain by subdomain and assembling the contributions as $\mathbf{r}_C = \sum_{i=1}^N \mathbf{R}_{Ci}^T \Psi_i^T \mathbf{E}_i^T \widehat{\mathbf{r}}$. The coarse correction is defined as the solution to problem $\mathbf{K}_C\,\mathbf{z}_C = \mathbf{r}_C$. Both corrections are finally added together and averaged on the interface by matrices $\mathbf{E}_i$ to produce the preconditioned residual $\widehat{\mathbf{z}} = \sum_{i=1}^N \mathbf{E}_i\,(\Psi_i \mathbf{R}_{Ci}\mathbf{z}_C + \mathbf{z}_i)$.

In the three-level BDDC method [8], the matrix $\mathbf{K}_C$ is not constructed on the second level. Instead, subdomains from the basic (first) level are grouped into subdomains on the next (second) level in the same way as elements of the original mesh are grouped into subdomains of the first level. The whole procedure described in this section is now repeated for the second level and thus the final coarse problem represents the third level. Obviously, this can be repeated again in the multilevel BDDC method. The only important difference between the first and the higher levels is the additional *interior pre-correction* and *post-correction* applied on higher levels in order to approximate the whole vector of coarse solution on the lower level.

Algorithm 2 Set-up of BDDC preconditioner with L levels

for level $\ell = 1, \ldots, L-1$ **do**
  if $\ell > 1$ **then**
    build *pseudo-mesh*: subdomains → 'elements'; corners + edges + faces → 'nodes'
  end if
  divide *pseudo-mesh* into subdomains (by METIS for $\ell > 1$, or by ParMETIS for $\ell = 1$)
  classify interface into *faces*, *edges*, *vertices*
  select *corners* (using face-based algorithm from [6])
  assemble matrices of subdomains $\mathbf{K}_i^\ell$ (use MPI to collect them on assigned cores)
  prepare interior correction – factorize interior block of $\mathbf{K}_i^\ell$ (serial MUMPS)
  factorize the matrices of local saddle point problems (1) (serial MUMPS)
  find *coarse basis functions* Ψ_i^ℓ and *coarse matrices* $\mathbf{K}_{Ci}^\ell = -\Lambda_i^\ell$ from (1) (serial MUMPS)
end for
factorize *global coarse matrix* $\mathbf{K}_C^{L-1} = \sum_{i=1}^{N_{L-1}} (\mathbf{R}_{Ci}^{L-1})^T \mathbf{K}_{Ci}^{L-1} \mathbf{R}_{Ci}^{L-1}$ (parallel MUMPS)

According to [4], the condition number of the operator preconditioned by multilevel BDDC with L levels satisfies $\kappa(\mathbf{M}_{BDDC}\widehat{\mathbf{S}}) \leq \prod_{\ell=1}^{L-1} C_\ell \left(1 + \log \frac{H_\ell}{H_{\ell-1}}\right)^2$, where H_ℓ is the characteristic size of subdomain on level ℓ, and $H_0 \equiv h$ is the characteristic size of element. Index ℓ is used here and throughout the next section to denote particular level. Due to the product present in this bound, each additional level worsens the mathematical efficiency of the multilevel preconditioner. The proof of the condition number bound as well as details of the algorithm of multilevel BDDC can be found in [4].

3 Parallel Implementation

Our implementation of the multilevel BDDC method has been recently released as an open-source solver library BDDCML.[1] It is written in Fortran 95 programming language and parallelized by MPI. The solver relies on the sparse direct solver MUMPS—a serial instance is used for each subdomain problem and a parallel instance is called for the final coarse problem. The solver supports assignment of several subdomains to each processor, since it is often useful to create divisions independently of number of available processors. A division of the mesh into subdomains on the first level is either provided to the solver by user's application or created internally by ParMETIS. The METIS package is currently used for this purpose on higher levels.

Similarly to other related preconditioners, we first need to set-up the multilevel BDDC preconditioner, which is then applied in each iteration of the Krylov subspace method. Details of the set-up are given in Algorithm 2, while key operations of each application are summarized in Algorithm 3. In these descriptions, we provide comments on how the steps are implemented in BDDCML in parentheses.

[1] http://www.math.cas.cz/sistek/software/bddcml.html

Algorithm 3 Application of BDDC preconditioner with L levels

for level $\ell = 1, \ldots, L-1$ **do**
  if $\ell > 1$ **then**
    $\widehat{r}^{\ell} \leftarrow \mathbf{r}_C^{\ell-1}$
    compute *interior pre-correction* of residual $\widehat{r}^{\ell}$ (serial MUMPS)
  end if
  distribute residual among subdomains $\mathbf{r}_i^{\ell} = (\mathbf{E}_i^{\ell})^T \widehat{r}^{\ell}$
  determine subdomain corrections $\mathbf{z}_i^{\ell}$ from (2) (serial MUMPS)
  construct coarse residual $\mathbf{r}_C^{\ell} = \sum_{i=1}^{N_\ell} (\mathbf{R}_{Ci}^{\ell})^T (\Psi_i^{\ell})^T (\mathbf{E}_i^{\ell})^T \widehat{r}^{\ell}$ (collective MPI)
end for
solve the coarse problem $\mathbf{K}_C^{L-1} \mathbf{z}_C^{L-1} = \mathbf{r}_C^{L-1}$ (parallel MUMPS)
for level $\ell = L-1, \ldots, 1$ **do**
  if $\ell < L-1$ **then**
    $\mathbf{z}_C^{\ell} \leftarrow \widehat{z}^{\ell+1}$
  end if
  combine coarse correction and subdomain corrections $\widehat{z}^{\ell} = \sum_{i=1}^{N_\ell} \mathbf{E}_i^{\ell} \left(\Psi_i^{\ell} \mathbf{R}_{Ci}^{\ell} \mathbf{z}_C^{\ell} + \mathbf{z}_i^{\ell} \right)$
  if $\ell > 1$ **then**
    apply *interior post-correction* to $\widehat{z}^{\ell}$ (serial MUMPS)
  end if
end for

4 Numerical Results

The first example corresponds to a problem of mechanical analysis of a cubic sample of geocomposite and was analyzed in [1]. The length of the edge of the cube is 75 mm. The cube comprises five distinct materials identified by means of computer tomography (Fig. 1 left), which causes anisotropic response of the cube even for simple axial stretching in z direction (Fig. 1 right). The problem is discretized using unstructured grid of about 12 million linear tetrahedral elements, resulting in approximately six million unknowns. The mesh was divided into 1,024, 128, and 16 subdomains on the first, second and third level, respectively, and the respective coarse problems (using corners and arithmetic averages on all edges and faces) contain 86,094, 11,265, and 612 unknowns.

Table 1 summarizes the efficiency of the multilevel preconditioner by means of the resulting condition number (estimated from the tridiagonal matrix generated during iterations of preconditioned conjugate gradient (PCG) method) and number of iterations. The iterations were stopped when the relative residual $\|\widehat{\mathbf{r}}\|/\|\widehat{\mathbf{g}}\|$ decreased bellow 10^{-6}. This table confirms the predicted worsening of the condition number with each additional level expected from the condition number bound.

Table 2 contains a strong scaling test using different number of levels. We differentiate the time spent on set-up and in PCG. All these computations were performed on the IBM SP6 computer at CINECA, Bologna. The computer is based on IBM Power6 4.7 GHz processors with 4 GB of RAM per core.

We can conclude from Table 2 that while adding levels seems not to be feasible for small number of cores (the computational time stagnates or even grows),

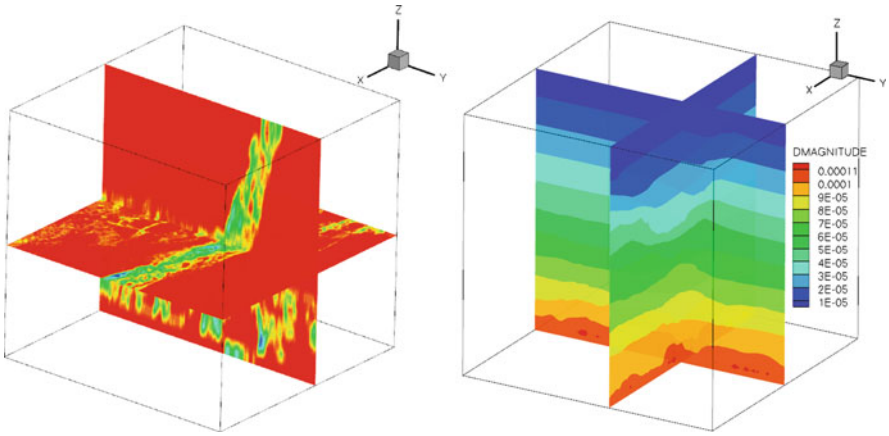


Fig. 1 Geocomposite problem: slices through material distribution (*left*) and displacement field (*right*)

Table 1 Condition number and number of iterations for different number of levels

Num. of levels	Num. of subs.	Cond. num.	Num. of PCG its
2	1024/1	50	46
3	1024/128/1	79	56
4	1024/128/16/1	568	131

Table 2 Strong scaling for geocomposite problem using two, three, and four levels

Number of processors	64	128	256	512	1,024
2 levels					
BDDC set-up time (s)	61.0	37.7	25.7	23.2	39.5
PCG time (s)	22.3	19.9	27.8	44.9	97.5
3 levels					
BDDC set-up time (s)	49.5	29.0	18.4	12.6	11.0
PCG time (s)	28.5	22.6	16.7	14.7	13.2
4 levels					
BDDC set-up time (s)	49.4	28.6	17.8	12.3	9.1
PCG time (s)	60.6	33.2	21.2	15.4	11.8

it improves the scaling on many cores. The minimal overall solution time is achieved for four levels and largest number of cores, despite the largest number of required iterations.

Our second example is a problem of Stokes flow in a 3D lid driven cavity. We use the set-up suggested in [9]: zero velocity is prescribed on all faces of the $[0, 1]^3$ cube except the face for z = 1, where unit velocity vector $\mathbf{u} = [1/\sqrt{3}, \sqrt{2/3}, 0]$ is

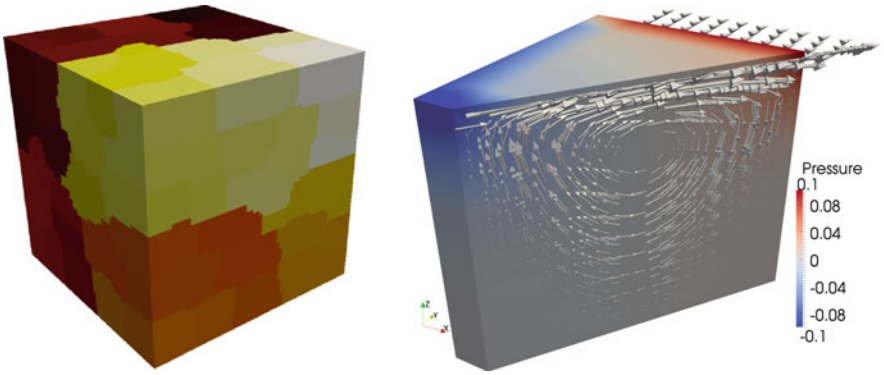


Fig. 2 Stokes flow in lid driven cavity: example of division into 64 subdomains (*left*), pressure contours and velocity vectors in the cut in direction of the prescribed velocity $[1/\sqrt{3}, \sqrt{2/3}, 0]$ (*right*)

Table 3 Weak scaling for Stokes flow in the cavity: BDDC using two and three levels

			BDDC (2 levels)			BDDC (3 levels)		
# elms.	# dofs. (M)	# cores	Divisions	# its.	Time (s)	Divisions	# its.	Time (s)
40^3	1.7	32	32/1	18	122	32/4/1	22	126
50^3	3.2	64	64/1	19	132	64/8/1	25	205
64^3	6.7	128	128/1	21	186	128/16/1	30	194
80^3	13.1	256	256/1	21	178	256/32/1	36	201
100^3	25.4	512	512/1	20	205	512/64/1	35	211

prescribed. We have used this test case also in the recent paper [5], but we have not presented parallel results there.

The problem is uniformly discretized using hexahedral Taylor–Hood finite elements. Computational mesh was divided into irregular partitions using the METIS graph partitioner (see Fig. 2 for an example). A plot of pressure inside the cavity and velocity vectors is given in Fig. 2 (right).

Table 3 summarizes a weak scaling test for this problem. The sequence of problems ranging from 1.7 million unknowns to 25.4 million unknowns are distributed among processors such that the size of local problems is kept approximately constant around 50,000 unknowns. We present results by the BDDC preconditioner using two and three levels combined with BiCGstab method. We report numbers of iterations and required overall computational times. The computational times were obtained on Darwin supercomputer of the University of Cambridge, using Intel Xeon 5100 3.0 GHz processors with 2 GB of RAM per core.

We can see in Table 3 that number of BiCGstab iterations remains almost constant for the two-level method, while mildly growing when using three levels, being again larger for the latter. The computational time slightly grows with problem size for BDDC, both using two and three levels, but this growth is rather acceptable.

More importantly, we can also see that for this case the benefit of using an additional level is slightly outweighed by the overhead of the additional iterations, and so the computational time is not improved by using three levels.

5 Conclusion

We have presented a parallel open-source implementation of the multilevel BDDC method. The two-level algorithm has scalability issues related to the coarse problem solution, mainly in the part of iterations. It can be noted that for the tested cases, it has not been the size of the coarse problem, but rather its fragmentation among too many cores which causes these issues. From our experiments, it appears that the multilevel preconditioner tends to scale better in both parts – set-up and Krylov subspace iterations. While the better scalability is able to translate into much faster solution for some cases, the extra overhead can also just cancel out the savings for other cases. It is therefore important to choose appropriate number of levels for a particular problem. We expect that advantages of the multilevel approach would pronounce further for problems divided into many (tens of thousands) of subdomains. Such challenging problems will likely become common in near future and will provide valuable feedback for further research in this field.

Acknowledgements We are grateful to Prof. Blaheta and Dr. Starý (Institute of Geonics AS CR) for providing the geocomposite problem. We are also grateful to Dr. Cirak (University of Cambridge) for providing computer time on Darwin. This work was supported by Ministry of Education, Youth and Sports of the Czech Republic under research project LH11004, by Czech Science Foundation under project 106/08/0403, by the Academy of Sciences of the CR through RVO: 67985840, by grant IAA100760702 of the Grant Agency of AS CR, by DOE through an ASCR grant, and by National Science Foundation under grant DMS-0713876, The research was started during two visits of Jakub Šístek at the University of Colorado Denver and some parts of the work have been performed under the HPC-Europa2 project with the support of the European Commission.

References

1. Blaheta, R., Jakl, O., Starý, J., Krečmer, K.: The Schwarz domain decomposition method for analysis of geocomposites. In: B. Topping et al. (eds.) Proceedings of the Twelfth International Conference on Civil, Structural and Environmental Engineering Computing. Civil-Comp Press, Stirlingshire, Scotland (2009)
2. Dohrmann, C.R.: A preconditioner for substructuring based on constrained energy minimization. SIAM J. Sci. Comput. **25**(1), 246–258 (2003)
3. Klawonn, A., Rheinbach, O.: Highly scalable parallel domain decomposition methods with an application to biomechanics. ZAMM Z. Angew. Math. Mech. **90**(1), 5–32 (2010)
4. Mandel, J., Sousedík, B., Dohrmann, C.R.: Multispace and multilevel BDDC. Computing **83** (2–3), 55–85 (2008)

5. Šístek, J., Sousedík, B., Burda, P., Mandel, J., Novotný, J.: Application of the parallel BDDC preconditioner to the Stokes flow. Comput. & Fluids **46**, 429–435 (2011)
6. Šístek, J., Čertíková, M., Burda, P., Novotný, J.: Face-based selection of corners in 3D substructuring. Math. Comput. Simulation **82**(10), 1799–1811 (2012)
7. Toselli, A., Widlund, O.B.: Domain Decomposition Methods—Algorithms and Theory, *Springer Series in Computational Mathematics*, vol. 34. Springer-Verlag, Berlin (2005)
8. Tu, X.: Three-level BDDC in three dimensions. SIAM J. Sci. Comput. **29**(4), 1759–1780 (2007)
9. Wathen, A.J., Loghin, D., Kay, D.A., Elman, H.C., Silvester, D.J.: A new preconditioner for the Oseen equations. In: F. Brezzi et al. (eds.) Numerical mathematics and advanced applications, pp. 979–988. Springer-Verlag Italia, Milano (2003). Proceedings of ENUMATH 2001

Forecasting Production in an Oil Reservoir Simulation and Its Challenges

V. Ginting, F. Pereira, and A. Rahunanthan

Abstract A Bayesian approach for uncertainty quantification of oil reservoir parameters in forecasting the production is straightforward in principle. However, the complexity of flow simulators and the nature of the inverse problem at hand present an ongoing practical challenges to addressing uncertainty in all subsurface parameters. In this paper, we focus on two important subsurface parameters, permeability and porosity, and discuss quantifying uncertainty in those parameters.

1 Introduction

The problems of conditioning simulator forecasting to data and quantifying uncertainty fit naturally into a Bayesian statistical framework in which mathematically and computationally numerical models of subsurface fluid flow are extensively used. The main aim is to quantify uncertainty in subsurface parameters for which our focus is directed toward permeability and porosity of the reservoir.

We propose a forecasting technique that is suitable for oil reservoir simulation through rigorous exploration of statistical distributions of permeability and porosity conditioned to partial production curves. The exploration is done using the Markov chain Monte Carlo algorithm (MCMC). A set of accepted fields is collected and used to predict the rest of the production curves. Two approaches are implemented: (1) use the accepted fields in the simulation to get a set of production curves that include the future prediction, and (2) evaluate the statistics of the accepted fields and

V. Ginting (✉) · A. Rahunanthan
Department of Mathematics, University of Wyoming, WY 82071, USA
e-mail: vginting@uwyo.edu; rahu@uwyo.edu

F. Pereira
Department of Mathematics and School of Energy Resources, University of Wyoming, WY 82071, USA
e-mail: lpereira@uwyo.edu

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_73,

use it in the simulation run to get a single production curve. Obviously the former requires a number of post-processing computations while the latter needs only one.

This paper is organized as follows. We discuss the physical and mathematical modeling of the problem at hand in Sect. 2. The parametrization of uncertainty using Karhunen-Loève expansion for unknown permeability and porosity fields is discussed in Sect. 3. In Sect. 4, we discuss a Bayesian approach for quantifying uncertainty in both permeability and porosity fields using an MCMC algorithm. In Sect. 5, we present simulation studies for the forecasting problem in an oil reservoir. Section 6 contains our conclusions.

2 Physical and Mathematical Modeling

The dynamics of the motion of fluids in a heterogeneous reservoir Ω can be categorized as an immiscible two-phase system with water and oil (denoted by w and o, respectively) that is incompressible. To simplify the model, capillary pressure and gravity are not included, and the two fluids fill the pore space. The reservoir is equipped with an injection well from which water is discharged to displace the trapped oil towards the production wells. The injection well is located in one of the corners. We have two production wells: one well is situated along the diagonal, opposite to the injection well, and the other is situated at the center of a side which is one of the two sides that enclose the production well at the corner (see Fig. 2). The wells are modeled through appropriate boundary conditions. The governing equations of flow and transport are

$$\nabla \cdot \boldsymbol{v} = 0, \quad \text{where} \quad \boldsymbol{v} = -\lambda(s)k(\boldsymbol{x})\nabla p, \quad \text{and} \quad \phi(\boldsymbol{x})\frac{\partial s}{\partial t} + \nabla \cdot (f(s)\boldsymbol{v}) = 0, \tag{1}$$

where $\boldsymbol{v}$ is the Darcy velocity, s is the water saturation, k is the absolute permeability and ϕ is the porosity. The total mobility $\lambda(s)$ and the flux function $f(s)$ are respectively given by:

$$\lambda(s) = \frac{k_{rw}}{\mu_w} + \frac{k_{ro}}{\mu_o}, \quad f(s) = \frac{k_{rw}(s)/\mu_w}{\lambda(s)}, \tag{2}$$

where k_{rj}, $j = w, o$, is the relative permeability of the phase j [1].

The oil production is modeled by the so-called fractional flow or production curve. For each production boundary the fractional flow $F(t)$ is defined as the fraction of oil in the produced fluid, i.e.,

$$F(t) = 1 - \frac{\displaystyle\int_{\partial\Omega_{\text{out}}} v_n f(s)\, dl}{\displaystyle\int_{\partial\Omega_{\text{out}}} v_n\, dl}, \quad \text{with PVI} = \int_0^T {V_p}^{-1} \int_{\partial\Omega_{\text{out}}} v_n\, dl\, d\tau, \tag{3}$$

where $\partial\Omega_{\text{out}}$ denotes outflow boundary, v_n is normal velocity field and t is the dimensionless time measured in Pore Volume Injected, PVI for short (T denotes the time taken for injection of water). The system (1) is simulated using an efficient and reliable physics-based operator splitting technique (see for example [7, 8] for further discussion), which is implemented on GPU machines.

3 Parametrization of Uncertainty

In the current work, we employ the Karhunen-Loève expansion (KLE) [6] to reduce the potentially large dimension of the uncertainty space describing the permeability and porosity which is accomplished through appropriate parametrization inherent in the expansion (see for example [3–5] for similar applications).

A standard assumption in geostatistics is to model the permeability to follow a log-normal distribution [2], i.e., $\log\left[k(\boldsymbol{x},\omega)\right] = Y^k(\boldsymbol{x},\omega)$, where $\boldsymbol{x} \in \Omega \subset \mathbf{R}^2$, and ω is a random element in a probability space, and $Y^k(\boldsymbol{x},\omega)$ is a field possessing a Gaussian distribution and a covariance function

$$R(\boldsymbol{x}_1,\boldsymbol{x}_2) = \sigma_Y^2 \exp\left(-\frac{1}{2}|\boldsymbol{L}^{-1}(\boldsymbol{x}_1-\boldsymbol{x}_2)|^2\right), \tag{4}$$

where, $\boldsymbol{L} = [L_x \; L_y]$ with $L_x = [L_{xx} \; L_{yx}]^\top$ and $L_y = [L_{xy} \; L_{yy}]^\top$ with the correlation lengths L_{ij}. The series representation of $Y^k(\boldsymbol{x},\omega)$ is

$$Y^k(\boldsymbol{x},\omega) = \sum_{i=1}^{\infty} Y_i^k(\omega)\varphi_i(\boldsymbol{x}), \quad \text{with} \quad Y_i^k(\omega) = \int_\Omega Y^k(\boldsymbol{x},\omega)\varphi_i(\boldsymbol{x})d\boldsymbol{x} \tag{5}$$

being functions of a random variable, and φ_i a set of basis functions satisfying

$$\int_\Omega R(\boldsymbol{x}_1,\boldsymbol{x}_2)\varphi_i(\boldsymbol{x}_2)d\boldsymbol{x}_2 = \lambda_i\varphi_i(\boldsymbol{x}_1), \;\; i = 1,2,\ldots, \tag{6}$$

that makes Y_i^k uncorrelated, and $\lambda_i = E[(Y_i^k)^2] > 0$. Denoting $\theta_i^k = Y_i^k/\sqrt{\lambda_i}$, then θ_i^k satisfies $E(\theta_i^k) = 0$ and $E(\theta_i^k\theta_j^k) = \delta_{ij}$, and thus

$$Y^k(\boldsymbol{x},\omega) = \sum_{i=1}^{\infty}\sqrt{\lambda_i}\theta_i^k(\omega)\varphi_i(\boldsymbol{x}) \simeq \sum_{i=1}^{N_k}\sqrt{\lambda_i}\theta_i^k\varphi_i(\boldsymbol{x}). \tag{7}$$

We assume that eigenvalues $\{\lambda_i\}_{i=1}^{\infty}$ are ordered so that $\lambda_1 \geq \lambda_2 \geq \cdots$. The basis functions $\varphi_i(\boldsymbol{x})$ in (6) are deterministic and resolve the spatial dependence of the permeability field and in particular its correlation structure. The uncertainty

is represented by the scalar random variables θ_i^k. In general, we only need to keep the leading order terms (quantified by the magnitude of λ_i) and still capture most of the energy of the stochastic process $Y^k(\boldsymbol{x}, \omega)$.

With respect to the porosity field, we make use of the standard assumption that the porosity exhibits a similar spatial correlation structure to the permeability. In turn, this allows us to employ (7). The dependence of porosity to the expansion is expressed as

$$\phi(\boldsymbol{x}) = \frac{\phi_{\min} + \phi_{\max} e^{Y^\phi}}{1 + e^{Y^\phi}}, \quad \phi_{\min} \text{ and } \phi_{\max} \in (0, 1), \tag{8}$$

where Y^ϕ is KLE for porosity as in (7), and $\phi_{\min}$ and $\phi_{\max}$ are the lower and upper limits of the porosity of the reservoir.

4 Bayesian Inference

As alluded to earlier, we want to sample the permeability and porosity fields conditioned on the available fractional flow data F_m. This is translated into sampling from the conditional distribution $P(\boldsymbol{\psi}|F_m)$, where $\boldsymbol{\psi} = [\boldsymbol{\theta}^k \ \boldsymbol{\theta}^\phi]$ with $\boldsymbol{\theta}^k$ and $\boldsymbol{\theta}^\phi$ vectors containing the random coefficients in the KLE. According to Bayes' theorem this distribution satisfies the proportionality relation

$$P(\boldsymbol{\psi}|F_m) \propto P(F_m|\boldsymbol{\psi}) P(\boldsymbol{\psi}), \tag{9}$$

where $P(F_m|\boldsymbol{\psi})$ represents the likelihood function (that requires the forward solution of the two-phase flow) and $P(\boldsymbol{\psi})$ is the prior distribution of $\boldsymbol{\psi}$. The normalizing constant in this expression is not important, because we use an iterative updating procedure. Although a more general error model can be used in the simulations, we assume that the likelihood function follows a Gaussian distribution. i.e.,

$$P(F_m|\boldsymbol{\psi}) \propto \exp\Big(-(F_m - F_{\boldsymbol{\psi}})^\top \Sigma (F_m - F_{\boldsymbol{\psi}}) \Big), \tag{10}$$

where $F_{\boldsymbol{\psi}}$ is the simulated fractional flow curve that is obtained by solving the forward problem with known permeability k and porosity ϕ, in other words with known $\boldsymbol{\psi}$, and Σ is the covariance matrix representing the measurement errors. We take $\Sigma = \boldsymbol{I}/2\sigma_F^2$, where $\boldsymbol{I}$ is the identity matrix and σ_F^2 is the precision associated with the measurement F_m and numerical solution $F_{\boldsymbol{\psi}}$.

In practice, both porosity and permeability might be dependent on each other, and that is the reason we take the same KL expansion structure for both fields. However, when we make a proposal, we assume that $\boldsymbol{\theta}^k$ and $\boldsymbol{\theta}^\phi$ are independent of each other and thus avoiding ad-hoc use of correlation between porosity and permeability.

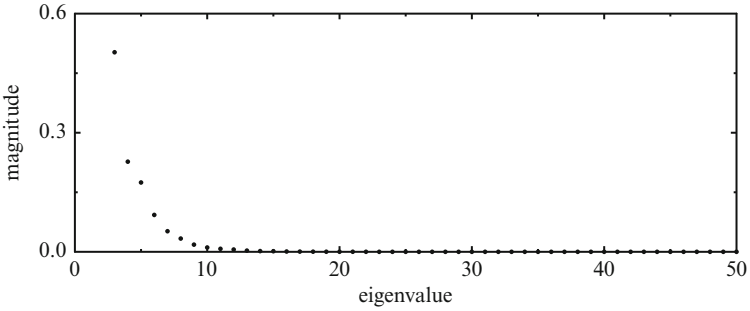


Fig. 1 Eigenvalues of the KLE for the Gaussian covariance with $L_x = L_y = 0.42$, $L_{xy} = L_{yx} = 0.65$ and $\sigma_Y^2 = 4$

We use the Metropolis-Hasting MCMC to sample from the posterior distribution. At each iteration, $\boldsymbol{\psi}_p = [\boldsymbol{\theta}_p^k \ \boldsymbol{\theta}_p^\phi]$ is proposed using an instrumental distribution $q(\boldsymbol{\psi}_p|\boldsymbol{\psi})$, where $\boldsymbol{\psi}$ represents the previously accepted state/parameters in the chain, and then the forward problem is solved to determine the acceptance probability,

$$\alpha(\boldsymbol{\psi}, \boldsymbol{\psi}_p) = \min\left(1, \frac{q(\boldsymbol{\psi}|\boldsymbol{\psi}_p)P(\boldsymbol{\psi}_p|F_m)}{q(\boldsymbol{\psi}_p|\boldsymbol{\psi})P(\boldsymbol{\psi}|F_m)}\right), \tag{11}$$

i.e., $\boldsymbol{\psi}_p$ is accepted with probability $\alpha(\boldsymbol{\psi}, \boldsymbol{\psi}_p)$.

5 Simulation Study

We now discuss the simulations of the two-phase flow problem in an oil reservoir as illustrated and present the associated numerical results. The relative permeability functions of water and oil take the form of s^2 and $(1-s)^2$, respectively, and the viscosity ratio between water and oil is 1:20. We assume that at $t = 0$, the reservoir is saturated by oil without any water, i.e., $s = 0$. The water is then injected at the injection well at the rate of one pore-volume every 5 years.

For the KLE that we use, Fig. 1 shows that the eigenvalues decay very fast, and it is enough to consider the first six eigenvalues in the KLE. Since we assume that the permeability and porosity share the same spatial structure, we share the same KLE structure for the permeability and porosity fields, with $N_k = N_\phi = 6$.

The forecasting of production curves in the oil reservoir consists of two steps: characterization and prediction. In the first step, we characterize the underlying field using the available data (in our case, the production curves until 0.25 PVI). As shown in Fig. 2, for three different sets of permeability and porosity distributions, we get similar saturation profiles and consequently similar production curves. Because of the nature of the inverse problem at hand, it is expected that we cannot recover

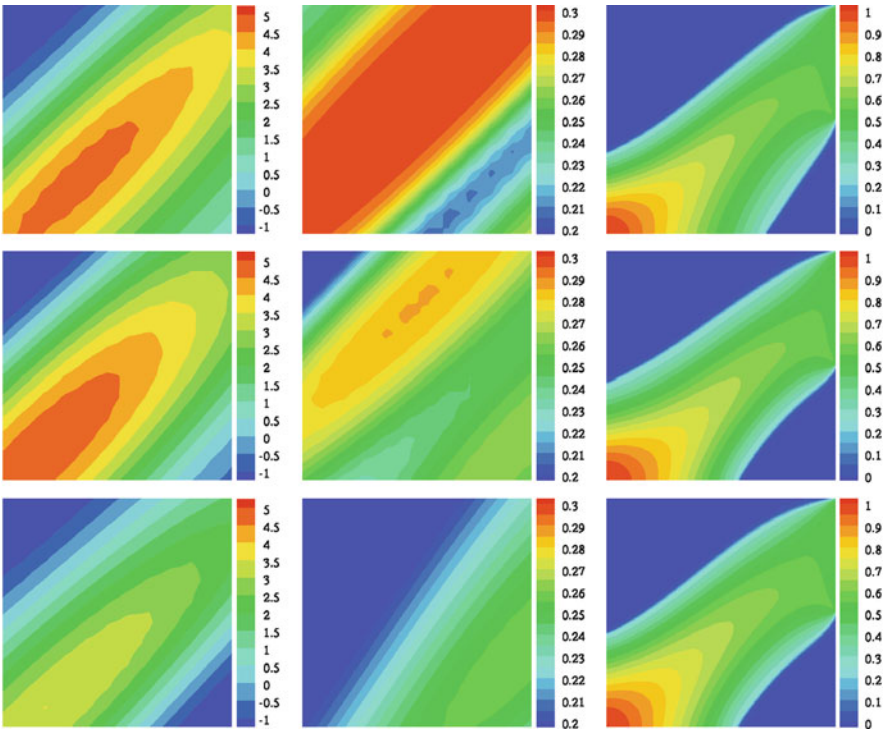


Fig. 2 A set of fields that give very similar fractional flow curves: *Left* to *right* the permeability and porosity distributions, and water saturation plot at $t = 1.25$ PVI, respectively

a unique profile for permeability and porosity nor is it necessary. In practice, it is more relevant to gather a set of probable profiles that has been rigorously selected by the Bayesian MCMC such as shown in Fig. 2.

In our investigation, there is an indication that the physics of the model dictates the dominance of the permeability over that of porosity in determining the flow pattern. This results in less information of the porosity that we are able to extract from the production curves. This makes the quantification of uncertainty for porosity more challenging than that of permeability. Figures 3 and 4 further illustrate this assessment. Figure 3 indicates that the resulting posterior distribution of permeability is considerably distinct from its prior information while that is not so much the case for the porosity as depicted in Fig. 4.

Next we look at the prediction of production curves. We assume that we have available production curves from the wells until 0.25 PVI. Therefore, we run the MCMC as described in Sect. 4 until 0.25 PVI with $\sigma_F^2 = 10^{-4}$ in the likelihood function (10). After selecting a set of accepted realizations of the parameters through the MCMC procedure, we run the forward problem until 1.25 PVI using those accepted realizations. We then aggregate the results of the forward problem.

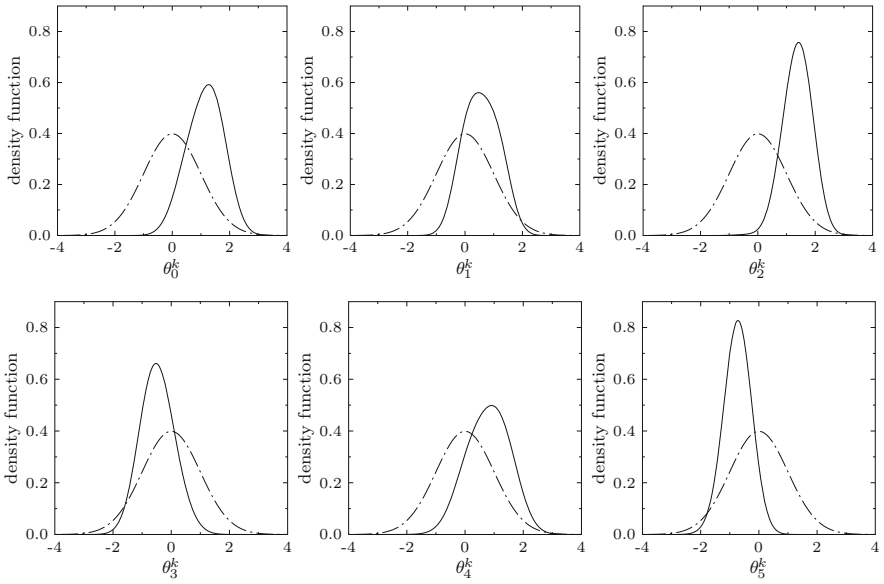


Fig. 3 Posterior exploration through the MCMC procedure for $\boldsymbol{\theta}^k$: *Solid line denotes* posterior and *dotted line denotes* the Gaussian prior

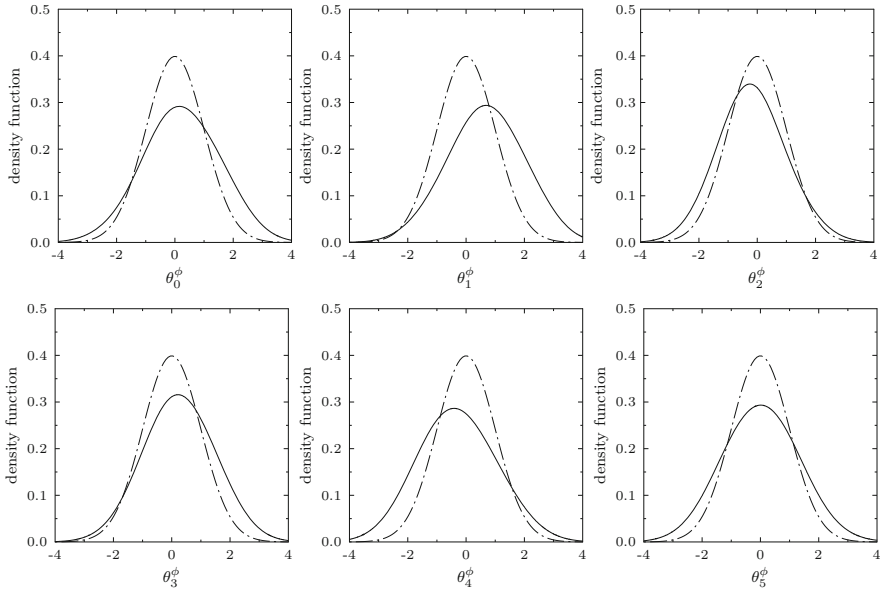


Fig. 4 Posterior exploration through the MCMC procedure for $\boldsymbol{\theta}^\phi$: *Solid line denotes* posterior and *dotted line denotes* the Gaussian prior

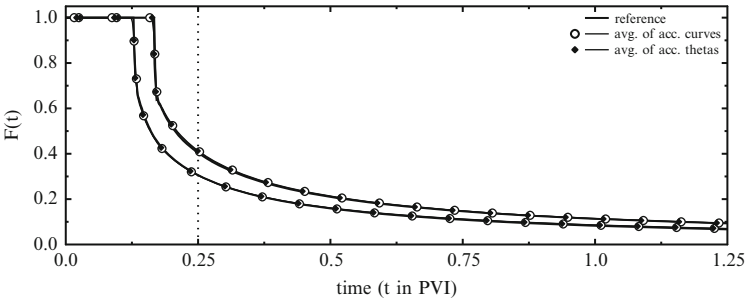


Fig. 5 A comparison of posterior mean production curves that includes recovery of the measurement and prediction. *Vertical line* marks $t = 0.25$ PVI beyond which prediction is performed

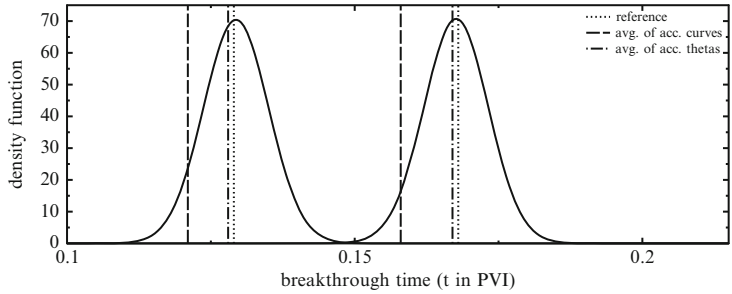


Fig. 6 Kernel density functions for breakthrough time of accepted production curves. *Vertical lines denote* the breakthrough time of the curves in Fig. 5

This average curve is referred to as the prediction of the production curve using *average of accepted production curves*. As other alternative, instead of running the forward problem several times, we calculate the statistics of the accepted parameters and run the forward problem only once. We refer this curve as the prediction of production curve using *average of accepted theta samples*. The plots Fig. 5 shows that the prediction in both approaches is very reliable. The vertical line separates the measured production curves and the predicted production curves. Although the former is more natural than the latter, the latter is not only reliable but also computationally efficient.

In Fig. 6, kernel density functions for breakthrough times of accepted production curves are shown. They give a hint that the simulated breakthrough time with the highest probability occurs precisely at the reference/measured breakthrough. Also, the figure illustrates that the breakthrough time obtained from using average of accepted theta samples is better than that of obtained from using average of accepted production curves. Since our main goal is to devise a reliable procedure for forecasting production, employing gauge of this type can serve as a good indicator for stopping the MCMC iterations.

6 Concluding Remarks

In this paper, we considered a forecasting problem in an oil reservoir. The partial production curves were used to characterize the reservoir in a Bayesian framework, and then, using accepted parameters from the characterization, we predicted the production curves of the production wells. The two approaches that we consider in this paper is very reliable with the latter requiring insignificant computational effort for the prediction after the MCMC procedure. In the future, we plan to investigate the possible application of the proposed framework to carbon geological sequestration.

Acknowledgements This work is partially supported by the grants from DOE (DE-FE0004832 and DE-SC0004982), the Center for Fundamentals of Subsurface Flow of the UW School of Energy Resources (WYDEQ49811GNTG, WYDEQ49811PER, WYDEQ49811FRTD), 2011 Clean Coal Technologies Research Program of the UW School of Energy Resources (1100 20352 2012), and from NSF (DMS-1016283).

References

1. Chen, Z., Huan, G., Ma, Y.: Computational methods for multiphase flows in porous media. SIAM, Philadelphia, PA (2006)
2. Dagan, G.: Flow and transport in porous formations. Springer-Verlag (1989)
3. Douglas, C., Efendiev, Y., Ewing, R., Ginting, V., Lazarov, R.: Dynamic data driven simulations in stochastic environments. Computing **77(4)**, 321–333 (2006)
4. Efendiev, Y., Datta-Gupta, A., Ginting, V., Ma, X., Mallick, B.: An efficient two-stage Markov chain Monte Carlo method for dynamic data integration. Water Resources Research **41(W12423)** (2005)
5. Ginting, V., Pereira, F., Presho, M., Wo, S.: Application of the two-stage Markov chain Monte Carlo method for characterization of fractured reservoirs using a surrogate flow model. Computational Geosciences **15(4)**, 691–707 (2011)
6. Loève, M.: Probability theory. Springer, Berlin (1977)
7. Pereira, F., Rahunanthan, A.: Numerical simulation of two-phase flows on a GPU. In: 9th International meeting on High Perfomance Computing for Computational Science (VECPAR '10). Berkeley, CA (2010)
8. Pereira, F., Rahunanthan, A.: A semi-discrete central scheme for the approximation of two-phase flows in three space dimensions. Mathematics and Computers in Simulation **81(10)**, 2296–2306 (2011)

Numerical Analysis for an Upscaled Model for Dissolution and Precipitation in Porous Media

K. Kumar, I.S. Pop, and F.A. Radu

Abstract In this paper, we discuss some numerical schemes for an upscaled (core scale) model describing the transport, precipitation and dissolution of solutes in a porous medium. We consider two weak formulations, conformal and mixed. We discuss the time discretization in both formulations and prove the convergence of the resulting schemes. A numerical study is presented for the mixed formulation.

1 Introduction

We consider reactive flow in a porous medium, where the ions/solutes are being transported through the combined process of convection and diffusion. Here a macroscale (upscaled) model is considered, meaning that no distinction is made between the solid grains and the pore space, and the equations are defined everywhere. Let $\Omega \subset \mathbb{R}^d$ $(d > 1)$ be the domain occupied by the porous medium, and assume Ω be open, connected, bounded and with Lipschitz boundary Γ. Further, let $T > 0$ be a finite time, and define

$$\Omega^T = (0, T] \times \Omega, \quad \text{and} \quad \Gamma^T = (0, T] \times \Gamma.$$

We consider here a simplified model, including only one mobile species. Denoting by v the concentration of the (immobile) precipitate, and by u the cation concentration, the model for the ion transport reduces to

K. Kumar (✉) · I.S. Pop
CASA, Technische Universiteit Eindhoven, PO Box 513, 5600 MB Eindhoven, The Netherlands
e-mail: K.Kumar@tue.nl; I.Pop@tue.nl

F.A. Radu
Institute of Mathematics, University of Bergen, Johannes Bruns gt. 12, Bergen, Norway
e-mail: Florin.Radu@math.uib.no

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_74,

$$\begin{cases} \partial_t(u+v) + \nabla \cdot (\mathbf{q}u - \nabla u) = 0, & \text{in } \Omega^T, \\ u = 0, & \text{on } \Gamma^T, \\ u = u_I, & \text{in } \Omega, \text{ for } t = 0, \end{cases} \tag{1}$$

and for the precipitate to

$$\begin{cases} \partial_t v = (r(u) - w), & \text{in } \Omega^T, \\ w \in H(v), & \text{in } \Omega^T, \\ v = v_I, & \text{in } \Omega, \text{ for } t = 0. \end{cases} \tag{2}$$

Here $\mathbf{q}$ stands for the Darcy fluid velocity. We assume that $\mathbf{q}$ is a known, divergence free velocity, i.e. $\nabla \cdot \mathbf{q} = 0$ in Ω.

For the ease of presentation we restrict to homogeneous Dirichlet boundary conditions. The initial data u_I and v_I are assumed non–negative and essentially bounded. Moreover, for simplicity we assume that $u_I \in H_0^1(\Omega)$.

All the quantities and variables in the above are assumed dimensionless. The diffusion is assumed 1, the extension to a positive definite diffusion tensor being straightforward. Further, we assume that the Damköhler number is scaled to 1, as well as an eventual factor in the time derivative of v in $(2)_1$, appearing in the transition form the pore scale to the core scale (see [8]). For the precipitation rate r we assume

(A1) $r(\cdot) : \mathbb{R} \to [0, \infty)$ is locally Lipschitz continuous in $\mathbb{R}$.
(A2) There exists a unique $u_* \geq 0$, such that

$$r(u) = \begin{cases} 0 \quad \text{for} \quad u \leq u_*, \\ \text{strictly increasing for} \quad u \geq u_* \quad \text{with} \quad r(\infty) = \infty. \end{cases} \tag{3}$$

The dissolution rate has a particular structure. It is assumed constant (1, by scaling) at some $(t, x) \in \Omega^T$ where the precipitate is present, i.e. if $v(t, x) > 0$. In the absence of the precipitate, the overall rate (precipitate minus dissolution) is either zero, if the solute present there is insufficient to produce a net precipitation gain, or positive. This can be summarized as

$$w \in H(v), \quad \text{where} \quad H(v) = \begin{cases} 0, & \text{if} \quad v < 0, \\ [0, 1] & \text{if} \quad v = 0, \\ 1 & \text{if} \quad v > 0. \end{cases} \tag{4}$$

Remark 1. In the setting above, a unique u^* exists for which $r(u^*) = 1$. If $u = u^*$ for all t and x, then the system is in equilibrium: no precipitation or dissolution occurs, since the precipitation rate is balanced by the dissolution rate regardless of the presence of absence of crystals. Then, as follows from [6–8, 10], for a.e. $(t, x) \in \Omega^T$ where $v = 0$, the dissolution rate satisfies

$$w = \begin{cases} r(u) \text{ if } & u < u^*, \\ 1 \quad \text{if} & u \geq u^*. \end{cases} \tag{5}$$

Remark 2. The upscaled model under discussion, proposed originally in [7] (see also [3–5]), is rigorously derived by homogenization techniques [8], starting from the pore scale counterpart in [6]. Next to the rigorous transition from the pore scale to the core scale, the homogenization procedure also provides the existence of solutions for the upscaled model.

We emphasize on the particularity of the present model, which is in the description of the dissolution and precipitation processes, involving a multi–valued dissolution rate. Clearly, classical solutions do not exist, except for some particular cases. Therefore we resort to defining appropriate weak solutions. These are defined in two ways. The first one is a conformal weak formulation (see e.g. [8]) and second is a mixed variational formulation (see [11–13] for similar problems). While the former is simpler and straightforward; the latter, separates the equation for the flux and retains the local mass conservation property.

2 The Conformal Weak Formulation

We start by defining the sets

$$\begin{aligned} \mathscr{U} &:= \{u \in L^2((0,T); H_0^1(\Omega)) : \partial_t u \in L^2((0,T); H^{-1}(\Omega))\}, \\ \mathscr{V} &:= \{v \in H^1((0,T); L^2(\Omega))\}, \\ \mathscr{W} &:= \{w \in L^\infty(\Omega^T), \ : 0 \leq w \leq 1\}. \end{aligned}$$

Throughout this work we use standard notations in the functional analysis. In particular, we mention that $(\cdot,\cdot)$ denotes the scalar product in $L^2(\Omega)$.

Definition 1. A triple $(u, v, w) \in \mathscr{U} \times \mathscr{V} \times \mathscr{W}$ is a weak solution of (1) and (2) if $(u_{|t=0}, v_{|t=0}) = (u_I, v_I)$, and for all $(\phi, \theta) \in (L^2(0, T, H^1_{0,\Gamma}(\Omega)), L^2(0, T; L^2(\Omega))$

$$\begin{aligned} (\partial_t u + \partial_t v, \phi) + (\nabla u, \nabla \phi) + (\mathbf{q}u, \nabla \phi) &= 0, \\ (\partial_t v, \theta) - \big(r(u) - H(v), \theta\big) &= 0, \\ w &\in H(v). \end{aligned} \tag{6}$$

We refer to [8] for the existence and uniqueness of a weak solution (see also Remark 1).

2.1 The Numerical Scheme

Here we provide a numerical scheme based on first order time stepping to approximate the solution in Definition 1. We construct a sequence of time discrete solutions,

which is then interpolated linearly in time. Letting the discretization parameter go to zero, we prove that the above approximation converges to the weak solution of (6). In addition to convergence, this also provides an alternative proof for the existence of solutions to (6).

Note the presence of a multi-valued rate in $(6)_3$, which impedes obtaining error estimates. Therefore we consider a regularized approximation of the original model (and pass later to the limit). With $\delta > 0$, define the regularized Heaviside function

$$H_\delta(v) = \begin{cases} 0, & \text{if} \quad v < 0, \\ \frac{v}{\delta} & \text{if} \quad 0 \le v \le \delta, \\ 1 & \text{if} \quad v > \delta. \end{cases} \tag{7}$$

Next, with $N \in \mathbb{N}$, $\tau = \frac{T}{N}$ and $t_n = n\tau, n = 1, \ldots, N$, we consider a uniform time stepping that is implicit in u and explicit in v. Starting with $u_\delta^0 = u_I$, $v_\delta^0 = v_I$, with $n \in \{1, \ldots, N\}$, the approximation (u_δ^n, v_δ^n) of $(u(t_n), v(t_n))$ solves

Problem 1 (P_δ^n)**.** Given $(u_\delta^{n-1}, v_\delta^{n-1}) \in \big(H_0^1(\Omega) \times L^2(\Omega)\big)$, find $(u_\delta^n, v_\delta^n) \in \big(H_0^1(\Omega) \times \in L^2(\Omega)\big)$ such that

$$\begin{aligned} \left(\frac{u_\delta^n - u_\delta^{n-1}}{\tau}, \phi\right) + (\nabla u_\delta^n, \nabla\phi) - (\mathbf{q}u_\delta^n, \nabla\phi) + \left(\frac{v_\delta^n - v_\delta^{n-1}}{\tau}, \phi\right) = 0 \\ \left(\frac{v_\delta^n - v_\delta^{n-1}}{\tau}, \theta\right) = \big(r(u_\delta^n) - H_\delta(v_\delta^{n-1}), \theta\big) \end{aligned} \tag{8}$$

for all $\phi \in H_0^1(\Omega), \theta \in L^2(\Omega)$. For completeness, we define

$$w_\delta^n := H_\delta(v_\delta^n).$$

For stability reasons, we choose $\delta = O(\tau^\alpha)$ for some $\alpha \in (0, 1)$ (see [2] for detailed arguments). By using $(8)_2$ in $(8)_1$ we get

$$\big(\frac{u_\delta^n - u_\delta^{n-1}}{\tau}, \phi\big) + (\nabla u_\delta^n, \nabla\phi) - (\mathbf{q}u_\delta^n, \nabla\phi) + \Big(r(u_\delta^n) - H_\delta(v_\delta^{n-1}), \phi\Big) = 0.$$

Standard monotonicity arguments provide the existence and uniqueness of a solution u_δ^n [9]. After computing u_δ^n, v_δ^n is obtained straightforwardly.

2.2 Convergence

By standard arguments (see e.g. [2]), a priori estimates can be obtained for the sequence of triples $(u_\delta^n, v_\delta^n, w_\delta^n)$ solving (8). Further, for any $t \in [t_{n-1}, t_n]$ we define

$$U^\tau(t) := u_\delta^n \frac{(t-t_{n-1})}{\tau} + u_\delta^{n-1} \frac{(t_n - t)}{\tau},$$
$$V^\tau(t) := v_\delta^n \frac{(t-t_{n-1})}{\tau} + v_\delta^{n-1} \frac{(t_n - t)}{\tau},$$
$$W^\tau(t) := H_\delta(V^\tau(t)),$$

In terms of the time–continuous approximation (U^τ, V^τ, W^τ), these estimates are stated in

Lemma 1. *There exist two constants $M, C > 0$, independent of τ and δ, such that*

$$0 \le U^\tau \le M, 0 \le V^\tau \le Me^{CT}, 0 \le W^\tau \le 1 \tag{9}$$
$$||U^\tau||^2 + ||V^\tau||^2 + ||\partial_t U^\tau||^2 + ||\nabla U^\tau||^2 + ||\partial_t V^\tau||^2 \le C. \tag{10}$$

The estimates in (9) should be understood in a. e. sense, whereas $\|\cdot\|$ is the usual norm in $L^2(\Omega^T)$. Clearly, $(U^\tau, V^\tau, W^\tau) \in \mathscr{U} \times \mathscr{V} \times L^\infty(\Omega)$. Since $\delta = O(\tau^\alpha)$ for some $\alpha \in (0,1)$, letting $\tau \searrow 0$ implies that both $\delta, \frac{\tau}{\delta} \searrow 0$. Since the estimates are uniform in τ and δ, we have

Lemma 2. *There exists $(u, v, w) \in \mathscr{U} \times \mathscr{V} \times \mathscr{W}$ s.t. along a sequence $\tau \searrow 0$*

(a) $U^\tau \rightharpoonup u$ *weakly in* $L^2((0,T); H_0^1(\Omega))$
(b) $\partial_t U^\tau \rightharpoonup \partial_t u$ *weakly in* $L^2((0,T); H^{-1}(\Omega))$
(c) $V^\tau \rightharpoonup v$ *weakly in* $L^2((0,T); L^2(\Omega))$
(d) $\partial_t V^\tau \rightharpoonup \partial_t v$ *weakly in* $L^2((0,T); L^2(\Omega))$
(e) $W^\tau \rightharpoonup w$ *weakly-star in* $L^\infty(\Omega^T)$.

To identify the limit triple (u, v, w) as the weak solution in Definition 1 stronger convergence properties are needed for V^τ, which can be obtained by translation estimates. This gives finally the existence of a weak solution to (6).

Theorem 1. *Along a sequence $\tau \searrow 0$, the triple (U^τ, V^τ, W^τ) converges to a limit (u, v, w) which is a weak solution in the sense of Definition 1.*

As proved in [8], the weak solution is unique, therefore the convergence holds along any sequence $\tau \searrow 0$.

3 The Mixed Variational Formulation

For introducing the mixed formulation of the original problem we define the set

$$\mathscr{S}_{\mathbf{Q}} := \{\mathbf{Q} \mid \mathbf{Q} \in L^2((0,T); H(div; \Omega))\},$$

where $H(div;\Omega)$ contains the vector-valued functions with divergence in L^2. A weak solution of (1)–(2) written in mixed form is introduced in

Definition 2. A quadruple $(u, \mathbf{Q}, v, w) \in (\mathscr{V} \times \mathscr{S}_{\mathbf{Q}} \times \mathscr{V} \times L^\infty((0,T)\times\Omega))$ with $u_{|t=0} = u_I$, $v_{|t=0} = v_I$ is a mixed weak solution of (1)–(2) if $w \in H(v)$ a.e. and for all $t \in (0,T)$ and $(\phi, \theta, \psi) \in H^1(\Omega) \times L^2(\Omega) \times H(div;\Omega)$ we have

$$\begin{aligned} \big(\partial_t u, \phi\big) + \big(\nabla\cdot\mathbf{Q}, \phi\big) + \big(\partial_t v, \phi\big) &= 0, \\ (\partial_t v, \theta) - \big(r(u) - H(v), \theta\big) &= 0, \\ \big(\mathbf{Q}, \psi\big) - (u, \nabla\cdot\psi) - (\mathbf{q}u, \psi) &= 0. \end{aligned} \tag{11}$$

3.1 The Numerical Scheme

As in Sect. 2.1, we consider the regularized problem obtained by replacing $H(\cdot)$ by $H_\delta(\cdot)$, and construct the sequence of time discrete approximations $(u_\delta^n, v_\delta^n, \mathbf{Q}_\delta^n, w_\delta^n)$ of $(u(t_n), \mathbf{Q}(t_n), v(t_n), w(t_n))$ by solving

Problem 2 ($P_\delta^{mvf,n}$). Given $(u_\delta^{n-1}, \mathbf{Q}_\delta^{n-1}, v_\delta^{n-1}) \in (L^2(\Omega), H(div;\Omega), L^2(\Omega))$ given, find $(u_\delta^n, \mathbf{Q}_\delta^n, v_\delta^n) \in (L^2(\Omega), H(div;\Omega), L^2(\Omega))$ such that

$$\begin{aligned} (u_\delta^n - u_\delta^{n-1}, \phi) + \tau(\nabla\cdot\mathbf{Q}_\delta^n, \phi) + (v_\delta^n - v_\delta^{n-1}, \phi) &= 0, \\ (v_\delta^n - v_\delta^{n-1}, \theta) - \tau(r(u_\delta^n), \theta) - \tau(H_\delta(v_\delta^{n-1}), \theta) &= 0, \\ (\mathbf{Q}_\delta^n, \psi) - (u_\delta^n, \nabla\cdot\psi) - (\mathbf{q}u_\delta^n, \psi) &= 0, \end{aligned} \tag{12}$$

for all $(\phi, \theta, \psi) \in H^1(\Omega) \times L^2(\Omega) \times H(div;\Omega)$. For completeness, define $w_\delta^n := H_\delta(v_\delta^n)$. As before, we choose $\delta = O(\tau^\alpha)$ for some $\alpha \in (0,1)$, and start with $u_\delta^0 = u_I$ and $v_\delta^0 = v_I$.

3.2 Convergence

Similar to the conformal case, convergence is proved by obtaining the a priori estimates for the time-discrete sequence $(u_\delta^n, v_\delta^n, \mathbf{Q}_\delta^n, w_\delta^n)$, defining the time-continuous approximations $(U^\tau, V^\tau, \mathbf{Q}^\tau, W^\tau)$ by linear interpolation, and using compactness arguments to identify a convergent sequence $(U^\tau, V^\tau, \mathbf{Q}^\tau, W^\tau)$ (as $\tau \searrow 0$) and identifying its limit as the mixed weak solution in Definition 2. The non-linearity in r leads to certain mathematical difficulties, solved by the strong L^2-convergence of U^τ. This is proved by using the bounds on the flux $\mathbf{Q}^\tau$, providing estimates for the spatial translation of U^τ. This is used further to enhance the convergence of V^τ, required further before passing to the limit in $(12)_2$. We conclude the above discussion with the main result of this section.

Theorem 2. *Along a sequence* $\tau \searrow 0$*, the quadruple* $(U^\tau, V^\tau, \mathbf{Q}^\tau, W^\tau)$ *converges to a limit* $(u, \mathbf{Q}, v, w)$ *which is a mixed weak solution of* (1)–(2)*.*

Remark 3. The convergence proof uses no estimates for the spatial gradients of $(U^\tau, V^\tau, \mathbf{Q}^\tau, W^\tau)$. Therefore the convergence results given here can be extended to the fully discrete case, obtained e.g. by employing a mixed finite element discretization in Problem $P_\delta^{mvf,n}$. We will address these issues in a forthcoming paper.

4 A Numerical Example

For testing purposes we consider a problem similar to (1) and (2), but including a right hand side in the first equation. This is chosen in such a way that the problem has an exact solution, which is used then to test the convergence of the mixed finite element scheme. Specifically, for $T = 1$ and $\Omega = (0, 5) \times (0, 1)$, and with $r(u) = [u]_+^2$ (where $[u]_+ := \max\{0, u\}$), we consider the problem

$$\begin{cases} \partial_t(u + v) + \nabla \cdot (\mathbf{q}u - \nabla u) = f, & \text{in } \Omega^T, \\ \partial_t v = (r(u) - w), & \text{on } \Omega^T, \\ w \in H(v), & \text{on } \Omega^T. \end{cases}$$

Here $\mathbf{q} = (1, 0)$ is a constant velocity, whereas

$$f(t, x, y) = \frac{1}{2} e^{x-t-5} \left(1 - e^{x-t-5}\right)^{-\frac{3}{2}} \left(1 - \frac{1}{2} e^{x-t-5}\right) - \begin{cases} 0, & \text{if } x < t, \\ e^{x-t-5}, & \text{if } x \geq t, \end{cases}$$

and the boundary and initial conditions are such that

$$u(t, x, y) = \left(1 - e^{x-t-5}\right)^{\frac{1}{2}} \quad \text{and} \quad v(t, x, y) = \begin{cases} 0, & \text{if } x < t, \\ \frac{e^{x-t}-1}{e^5} & \text{if } x \geq t, \end{cases}$$

$$\text{providing} \quad w(t, x, y) = \begin{cases} 1, & \text{if } x < t, \\ 1 - e^{x-t-5} & \text{if } x \geq t, \end{cases}$$

form a solution triple.

We consider the mixed finite element discretization of the problem above, based on the time stepping in Sect. 3 and the lowest order Raviart-Thomas elements RT_0. The numerical scheme was implemented in the software package *ug* [1]. The simulations are carried out for a constant mesh diameter h and time step τ, satisfying $\tau = h$. Further, we take $\delta = \sqrt{h}$ as regularizing parameter. We start with $h = 0.2$, and refine the mesh (and correspondingly τ and δ) four times successively by halving h up to $h = 0.0125$. We compute the errors for u and v in the L^2 norms,

Table 1 Convergence results for the mixed discretization, $h = \tau$ and $\delta = \sqrt{\tau}$

h	$\|u - U^\tau\|$	α	$\|v - V^\tau\|$	β
0.2	1.1700e−01		1.8409e−01	
0.1	6.414e−02	0.87	9.927e−02	0.89
0.05	3.396e−02	0.91	5.317e−02	0.90
0.025	1.726e−02	0.98	2.785e−02	0.93
0.0125	8.42e−03	1.03	1.420e−02	0.97

$$E_u^h = \|u - U^\tau\|_{L^2(\Omega^T)}, \quad \text{respectively} \quad E_v^h = \|v - V^\tau\|_{L^2(\Omega^T)}.$$

These are presented in Table 1. Although theoretically no error estimates could be given due to the particular character of the dissolution rate, Table 1 also includes an estimate of the convergence order, based on the reduction factor between two successive calculations:

$$\alpha = \log_2(E_u^h / E_u^{\frac{h}{2}}), \qquad \text{and} \qquad \beta = \log_2(E_v^h / E_v^{\frac{h}{2}}).$$

For this simple test case, the method converges linearly.

Acknowledgements The financial support of STW (project 07796) is gratefully acknowledged by K. Kumar.

References

1. P. Bastian, K. Birken, K. Johannsen, S. Lang, N. Neuss, H. Rentz-Reichert, and C. Wieners. UG – a flexible software toolbox for solving partial differential equations. *Comput. and Vis. in Science*, 1:27–40, 1997.
2. V. M. Devigne, I. S. Pop, C. J. van Duijn, and T. Clopeau. A numerical scheme for the pore-scale simulation of crystal dissolution and precipitation in porous media. *SIAM J. Numer. Anal.*, 46(2):895–919, 2008.
3. C. J. van Duijn and P. Knabner. Solute transport through porous media with slow adsorption. In *Free boundary problems: theory and applications, Vol. I (Irsee, 1987)*, volume 185 of *Pitman Res. Notes Math. Ser.*, pages 375–388. Longman Sci. Tech., Harlow, 1990.
4. C. J. van Duijn and P. Knabner. Solute transport in porous media with equilibrium and nonequilibrium multiple-site adsorption: travelling waves. *J. Reine Angew. Math.*, 415:1–49, 1991.
5. C. J. van Duijn and P. Knabner. Travelling wave behaviour of crystal dissolution in porous media flow. *European J. Appl. Math.*, 8(1):49–72, 1997.
6. C. J. van Duijn and I. S. Pop. Crystal dissolution and precipitation in porous media: pore scale analysis. *J. Reine Angew. Math.*, 577:171–211, 2004.
7. P. Knabner, C. J. van Duijn, and S. Hengst. An analysis of crystal dissolution fronts in flows through porous media. part 1: Compatible boundary conditions. *Adv. Water Resour.*, 18:171–185, 1995.
8. K. Kumar, M. Neuss-Radu, and I. S. Pop. Homogenization of crystal precipitation and dissolution model in a porous medium. *In Preparation*.

9. O. A. Ladyzhenskaya and N. N. Ural′tseva. *Linear and quasilinear elliptic equations*. Translated from the Russian by Scripta Technica, Inc. Translation editor: Leon Ehrenpreis. Academic Press, New York, 1968.
10. T. L. van Noorden, I. S. Pop, and M. Röger. Crystal dissolution and precipitation in porous media: L^1-contraction and uniqueness. *Discrete Contin. Dyn. Syst.*, (Dynamical Systems and Differential Equations. Proceedings of the 6th AIMS International Conference, suppl.):1013–1020, 2007.
11. F. A. Radu and I. S. Pop. Newton method for reactive solute transport with equilibrium sorption in porous media. *J. Comput. Appl. Math.*, 234(7):2118–2127, 2010.
12. F. A. Radu, I. S. Pop, and S. Attinger. Analysis of an Euler implicit-mixed finite element scheme for reactive solute transport in porous media. *Numer. Methods Partial Differential Equations*, 26(2):320–344, 2010.
13. F. A. Radu, I. S. Pop, and P. Knabner. Error estimates for a mixed finite element discretization of some degenerate parabolic equations. *Numer. Math.*, 109(2):285–311, 2008.

A Variational Multiscale Method for Poisson's Equation in Mixed Form

M.G. Larson, A. Målqvist, and R. Söderlund

Abstract In this paper we present the adaptive variational multiscale method for solving the Poisson equation in mixed form. We use the method introduced in Larson and Målqvist (Comput Method Appl Mech Eng 196:2313–2324, 2007), and further analyzed and applied to mixed problems in Larson and Målqvist (Comput Method Appl Mech Eng 19:1017–1042, 2009), which is a general tool for solving linear partial differential equations with multiscale features in the coefficients. We extend the numerics in Larson and Målqvist (Comput Method Appl Mech Eng 19:1017–1042, 2009) from rectangular meshes to triangular meshes which allow for computation on more complicated domains. A new a posteriori error estimate is also included, which is used in an adaptive algorithm. We present a numerical example that shows the efficiency of incorporating a posteriori based adaptivity into the method.

M.G. Larson (✉)
Professor of Applied Mathematics, Department of Mathematics, Umeå University, SE-901-87 Umeå, Sweden
e-mail: mats.larson@math.umu.se

A. Målqvist
Assistant Professor, Department of Information Technology, Uppsala University, SE-751-05 Uppsala, Sweden
e-mail: axel.malqvist@it.uu.se

R. Söderlund
Research Assistant, Department of Mathematics, Umeå University, SE-901-87 Umeå, Sweden
e-mail: robert.soderlund@math.umu.se

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_75,

1 Introduction

Multiscale problems appear in many applications in engineering and sciences, for instance, composite materials, flow in porous media, fluid mechanics, and quantum physics. A common feature of multiscale problems is that they are very computationally challenging and often impossible to solve to an acceptable tolerance with standard methods using only one mesh. Thus multiscale methods are introduced, which uses both local and global information computed on different scales.

Multiscale methods have been developed in various ways the last 15 years. A common feature is that information from decoupled local fine scale equations are used to modify the coarse scale solution. Two early examples are the multiscale finite element method [1] and the variational multiscale method [2]. The adaptive variational multiscale method was first introduced by Larson and Målqvist in [3], which presents a posteriori error estimates that can be used in adaptive algorithms. That method is extended to mixed problems in [4], and further developed in e.g [5]. This paper is based on [4], but we extend the numerics to triangular meshes which allows for more complicated geometries. We also derive a new improved a posteriori error estimate. This error estimate is used in an adaptive algorithm, that automatically tunes the parameters of the method.

2 Preliminaries

We let $\Omega \subset \mathbf{R}^d$ be a domain with Lipschitz boundary $\partial\Omega$. We consider a coarse scale and a fine scale, both of which need to be discretized. We denote the coarse mesh by $\mathscr{K}_H$, with H_K we refer to the diameter of the elements in the coarse mesh, and we let $H = \max_K H_K$. The coarse mesh satisfies $\cup_{K\in\mathscr{K}_H} K = \Omega$ where all K are disjoint. The fine mesh however, is only defined on local subregions $\omega \subset \Omega$, since we wish to decouple the fine scale computations. The meshes will be nested so that all of those subregions are made up of coarse elements. We therefore introduce the following notation, $\mathscr{K}_H(\omega) = \{K \in \mathscr{K}_H : K \in \omega\}$ and with $\mathscr{K}_h(\omega)$ we refer to the set of fine scale elements $\{K\}$ such that $\cup_{K\in\mathscr{K}_h(\omega)} K = \omega$. Since the meshes are nested all $K \in \mathscr{K}_H(\omega)$ can be written as a union of elements in $\mathscr{K}_h(\omega)$. The diameter of the elements $K \in \mathscr{K}_h(\Omega)$ will be denoted h_K, and we let $h = max_K h_K$.

Next we define the function spaces $\mathscr{V} = \{\boldsymbol{v} \in H(div;\Omega) : \boldsymbol{n}\cdot\boldsymbol{v} = 0 \text{ on } \partial\Omega\}$ and $\mathscr{W} = L^2(\Omega)$, where $\boldsymbol{n}$ is the outward unit normal to $\partial\Omega$. The corresponding finite element spaces on the coarse scale will be denoted $\mathscr{V}_c$ and $\mathscr{W}_c$ respectively, and we let

$$\mathscr{V}_c = \mathscr{RT}_H, \tag{1a}$$

$$\mathscr{W}_c = \mathscr{P}_H, \tag{1b}$$

where $\mathscr{P}_H$ is the space of piecewise constants on the coarse mesh, and $\mathscr{RT}_H$ is the space of lowest order Raviart-Thomas elements on the coarse mesh. We want

an hierarchical split between the coarse and the fine scales and if we introduce the operators $\Pi_H : \mathscr{V} \to \mathscr{V}_c$ and $P_H : \mathscr{W} \to \mathscr{W}_c$, as the Raviart-Thomas interpolant and L^2-projection onto the coarse scale spaces respectively, we can define the fine scale spaces $\mathscr{V}_f, \mathscr{W}_f$ by

$$\mathscr{V}_f = \{\boldsymbol{v} \in \mathscr{V} : \Pi_H \boldsymbol{v} = \boldsymbol{0}\}, \tag{2a}$$

$$\mathscr{W}_f = \{v \in \mathscr{W} : P_H v = 0\}. \tag{2b}$$

We let $\mathscr{P}_h(\omega)$ and $\mathscr{RT}_h(\omega)$ be the piecewise constants, and the lowest order Raviart-Thomas elements, on the subgrid $\mathscr{K}_h(\omega)$. The fine scale finite element spaces can now be defined in the following way

$$\mathscr{P}_h^f(\omega) = \{v \in \mathscr{P}_h(\omega) : P_H v = 0\}, \tag{3a}$$

$$\mathscr{RT}_h^f(\omega) = \{\boldsymbol{v} \in \mathscr{RT}_h(\omega) : \Pi_H \boldsymbol{v} = \boldsymbol{0} \text{ and } \boldsymbol{n} \cdot \boldsymbol{v} = 0 \text{ on } \partial\omega\}, \tag{3b}$$

where, $\mathscr{P}_h^f(\omega)$ approximates $\mathscr{W}_f$ and $\mathscr{RT}_h^f(\omega)$ approximates $\mathscr{V}_f$ on the subregion ω.

The patches ω_i on which we define the fine scale finite element spaces, are defined in the following way.

Definition 1. We say that ω_i^1 is a 1-layer patch if $\omega_i^1 = \text{supp}(\boldsymbol{\phi}_i)$, where $\boldsymbol{\phi}_i$ is a coarse Raviart-Thomas basis function. Further we say that ω_i^n is an n-layer patch if,

$$\omega_i^n = \cup_{\{i:\text{supp}(\theta_i)\cap\omega_i^{n-1}\neq\emptyset\}}\text{supp}(\theta_i), \quad n = 2, 3, \ldots \tag{4}$$

where θ_i is a coarse scale continous piecewise linear nodal basis function. In the text we omit the superscript n.

Let us also introduce the fine scale projection operator $P_{h,\omega} : \mathscr{W} \to \mathscr{P}_h^f(\omega)$ and the Scott-Zhang interpolants $P_H^1 : \mathscr{V} \to \mathscr{P}_H^1$ and $P_{h,\omega,0}^1 : \mathscr{V} \to \mathscr{P}_{h,0}^1$, where the latter projects onto functions that are zero on the boundary.

3 A Variational Multiscale Decomposition of Poisson's Equation in Mixed Form

The equation we wish to solve is the Poisson equation in mixed form, which reads: find the pressure $u \in \mathscr{W}$ and the flux $\boldsymbol{\sigma} \in \mathscr{V}$ such that

$$\tfrac{1}{a}\boldsymbol{\sigma} = \nabla u, \qquad \text{in } \Omega, \tag{5a}$$

$$-\nabla \cdot \boldsymbol{\sigma} = f, \qquad \text{in } \Omega, \tag{5b}$$

$$n \cdot \boldsymbol{\sigma} = 0, \qquad \text{on } \partial\Omega, \tag{5c}$$

where $a \in L^\infty(\Omega)$ is the permeability satisfying $a \geq a_0 > 0$ for some constant a_0, and $f \in L^2(\Omega)$ is a given external force such that $\int_\Omega f dx = 0$. We use the variational multiscale framework, see [2], and let $u = u_c + u_f$, $\boldsymbol{\sigma} = \boldsymbol{\sigma}_c + \boldsymbol{\sigma}_f$, where $\boldsymbol{\sigma}_c \in \mathscr{V}_c$, $u \in \mathscr{W}_c$, $\boldsymbol{\sigma}_f \in \mathscr{V}_f$, and $u_f \in \mathscr{W}_f$. Let φ_k denote the piecewise constant basis functions on the coarse mesh, such that $\mathscr{W}_c = span(\{\varphi_k\}_{\mathscr{M}})$, where $\mathscr{M}$ is the set of coarse scale elements. Also let $\mathscr{N}$ be the set of coarse faces and ψ_i be the partition of unity defined by $\psi_i = \frac{1}{d+1}$ on $supp(\boldsymbol{\phi}_i)$ ($\boldsymbol{\phi}_i$ being the continuous piecewise linear nodal basis functions). As in [4], the multiscale finite element solution $\boldsymbol{\Sigma}$ corresponding to $\boldsymbol{\sigma}$, is given by $\boldsymbol{\Sigma} = \sum_{i \in \mathscr{N}} \Sigma_c^i(\boldsymbol{\phi}_i + \boldsymbol{\xi}_i) + \boldsymbol{\beta}$, where $\boldsymbol{\beta} = \sum_{i \in \mathscr{N}} \boldsymbol{\beta}_i$, and

$$(\tfrac{1}{a} \sum_{j \in \mathscr{N}} \Sigma_c^j \boldsymbol{\phi}_j, \boldsymbol{\phi}_i) + (\tfrac{1}{a} \sum_{j \in \mathscr{N}} \Sigma_c^j \boldsymbol{\xi}_j, \boldsymbol{\phi}_i) - (\sum_{k \in \mathscr{M}} U_c^k \varphi_k, \nabla \cdot \boldsymbol{\phi}_i) = -(\tfrac{1}{a}\boldsymbol{\beta}, \boldsymbol{\phi}_i), \tag{6a}$$

$$(\nabla \cdot \sum_{j \in \mathscr{N}} \Sigma_c^j \boldsymbol{\phi}_j, \varphi_k) = -(f, \varphi_k), \tag{6b}$$

$$\forall i \in \mathscr{N}, \forall k \in \mathscr{M},$$

$$(\tfrac{1}{a}\boldsymbol{\xi}_i, \boldsymbol{v}_f) + (\eta_i, \nabla \cdot \boldsymbol{v}_f) = -(\tfrac{1}{a}\boldsymbol{\phi}_i, \boldsymbol{v}_f), \qquad \forall \boldsymbol{v}_f \in \mathscr{RT}_h^f(\omega_i), \forall i \in \mathscr{N}, \tag{7a}$$

$$-(\nabla \cdot \boldsymbol{\xi}_i, w_f) = 0, \qquad \forall w_f \in \mathscr{P}_h^f(\omega_i), \forall i \in \mathscr{N}, \tag{7b}$$

$$(\tfrac{1}{a}\boldsymbol{\beta}_i, \boldsymbol{v}_f) + (\rho_i, \nabla \cdot \boldsymbol{v}_f) = 0, \qquad \forall \boldsymbol{v}_f \in \mathscr{RT}_h^f(\omega_i), \forall i \in \mathscr{N}, \tag{8a}$$

$$-(\nabla \cdot \boldsymbol{\beta}_i, w_f) = (f, \psi_i w_f), \qquad \forall w_f \in \mathscr{P}_h^f(\omega_i), \forall i \in \mathscr{N}. \tag{8b}$$

4 A Posteriori Error Estimate

In this section we present an a posteriori error estimate for the proposed multiscale method. For simplicity we assume two spatial dimensions and that a is piecewise constant. We follow ideas presented in [6], page 26–29. We start by presenting a technical Lemma.

Lemma 1. *Let $\mathscr{P}_h^1(\omega) \subset H^1(\omega)$ be the space of continuous piecewise linear functions and let $\mathscr{RT}_h(\omega)$ be the space of lowest order Raviart-Thomas finite elements on a given triangulation $\mathscr{K}$ of a domain $\omega \subset \Omega$. Further let $\nabla \times \phi = [\partial\phi/\partial y, -\partial\phi/\partial x]$, for any $\phi \in H^1(\omega)$. Then*

(i) $\nabla \times \phi_h \in \mathscr{RT}_h$ *for all* $\phi_h \in \mathscr{P}_h^1(\omega)$.
(ii) For any function ϕ_h *that vanishes on the boundary* $\partial\omega$ *we have that* $\boldsymbol{n} \cdot \nabla \times \phi_h = 0$ *on the boundary* $\partial\omega$, $\boldsymbol{n}$ *being the normal of the boundary* $\partial\omega$.

Proof. For (i) we refer to [6] and (ii) is easily seen since if ϕ_h vanishes on the boundary, the gradient $\nabla\phi_h = [\partial\phi_h/\partial x, \partial\phi_h/\partial y]$ must be parallell to $\boldsymbol{n}$ and thus the curl $\nabla \times \phi_h = [\partial\phi_h/\partial y, -\partial\phi_h/\partial x]$ must be orthogonal to $\boldsymbol{n}$.

We are now ready to present the main theorem.

Theorem 1. *We let* $\boldsymbol{\Sigma} = \sum_{i\in\mathscr{N}} \Sigma_c^i(\boldsymbol{\phi}_i + \boldsymbol{\xi}_i) + \boldsymbol{\beta}$ *be the multiscale approximation of* $\boldsymbol{\sigma}$ *and assume* $d = 2$. *It holds,*

$$\|\frac{1}{\sqrt{a}}(\boldsymbol{\sigma} - \boldsymbol{\Sigma})\|_{L^2(\Omega)}^2 \leq C \sum_{i\in\mathscr{N}} \sum_{K\in\mathscr{K}(\omega_i)} h_K^2 \|f\psi_i + \nabla\cdot(\Sigma_c^i(\boldsymbol{\phi}_i + \boldsymbol{\xi}_i) + \boldsymbol{\beta}_i)\|_{L^2(K)}^2 \tag{9}$$

$$+ \sum_{i\in\mathscr{N}} \sum_{K\in\mathscr{K}_h(\omega_i)} h_K \|[\boldsymbol{t}\cdot\frac{1}{a}(\Sigma_c^i(\boldsymbol{\phi}_i + \boldsymbol{\xi}_i) + \boldsymbol{\beta}_i)]\|_{L^2(\partial K)}^2$$

$$+ \sum_{i\in\mathscr{N}} H\|\tilde{\Sigma}^i\|_{L^2(\partial\omega_i)}^2$$

where $[\cdot]$ *denotes the jump across the boundary* ∂K, $\boldsymbol{t}$ *is the tangent vector to* ∂K, *and* $\tilde{\Sigma}^i \in \mathscr{P}_h^1(\partial\omega_i)$ *is defined for each* $i \in \mathscr{N}$ *as the solution to,*

$$(\tilde{\Sigma}^i, v)_{L^2(\partial\omega_i)} = (\frac{1}{a}(\Sigma_c^i(\boldsymbol{\phi}_i + \boldsymbol{\xi}_i) + \boldsymbol{\beta}_i), \nabla\times v)_{L^2(\omega_i)}, \qquad \text{for all } v \in \mathscr{P}_h^1(\omega_i). \tag{10}$$

Proof. We note that there exists functions $\phi \in H^1(\Omega)$ and $\gamma \in H_0^1(\Omega)$ such that, $\boldsymbol{v} = a\nabla\gamma + \nabla\times\phi$ for all $\boldsymbol{v} \in (L^2(\Omega))^2$ and furthermore

$$\|\nabla\gamma\|_{L^2(\Omega)} + \|\nabla\phi\|_{L^2(\Omega)} \leq C\|\boldsymbol{v}\|_{L^2(\Omega)}. \tag{11}$$

We let $\boldsymbol{\sigma} - \boldsymbol{\Sigma} = a\nabla\gamma + \nabla\times\phi$ and get,

$$\|\frac{1}{\sqrt{a}}(\boldsymbol{\sigma} - \boldsymbol{\Sigma})\|_{L^2(\Omega)}^2 = (\boldsymbol{\sigma} - \boldsymbol{\Sigma}, \nabla\gamma) + (\frac{1}{a}(\boldsymbol{\sigma} - \boldsymbol{\Sigma}), \nabla\times\phi) = \mathrm{I} + \mathrm{II}. \tag{12}$$

We treat the two terms separately.

We start with the first term and use Green's formula and the orthogonality given by the multiscale method, as well as interpolation estimates and equation (11), with a modified constant C depending on a.

$$\mathrm{I} = (\boldsymbol{\sigma} - \boldsymbol{\Sigma}, \nabla\gamma) = (-\nabla \cdot (\boldsymbol{\sigma} - \boldsymbol{\Sigma}), \gamma) \tag{13}$$

$$= (f + \nabla \cdot \boldsymbol{\Sigma}, \gamma - P_H\gamma) \tag{14}$$

$$= \sum_{i \in \mathcal{N}} (f\psi_i + \nabla \cdot (\Sigma_c^i(\boldsymbol{\phi}_i + \boldsymbol{\xi}_i) + \boldsymbol{\beta}_i), \gamma - P_H\gamma - P_{h,\omega_i}\gamma) \tag{15}$$

$$\leq C \left(\sum_{i \in \mathcal{N}} \sum_{K \in \mathcal{K}_h(\omega_i)} h_K^2 \| f\psi_i + \nabla \cdot (\Sigma_c^i(\boldsymbol{\phi}_i + \boldsymbol{\xi}_i) + \boldsymbol{\beta}_i) \|_{L^2(K)}^2 \right)^{1/2} \tag{16}$$

$$\| \frac{1}{\sqrt{a}} (\boldsymbol{\sigma} - \boldsymbol{\Sigma}) \|_{L^2(\Omega)},$$

Next we turn to the second term. We note that P^1 is the Scott-Zhang interpolant onto continuous piecewise linear functions and use Lemma 1 to subtract the curl of a continuous piecewise linear function,

$$\mathrm{II} = (\frac{1}{a}(\boldsymbol{\sigma} - \boldsymbol{\Sigma}), \nabla \times \phi) = -(\frac{1}{a}\boldsymbol{\Sigma}, \nabla \times \phi) \tag{17}$$

$$= -(\frac{1}{a}\boldsymbol{\Sigma}, \nabla \times (\phi - P_H^1\phi)) \tag{18}$$

$$= -\sum_{i \in \mathcal{N}} (\frac{1}{a}(\Sigma_c^i(\boldsymbol{\phi}_i + \boldsymbol{\xi}_i) + \boldsymbol{\beta}_i), \nabla \times (\phi - P_H^1\phi)) \tag{19}$$

$$= -\sum_{i \in \mathcal{N}} (\frac{1}{a}(\Sigma_c^i(\boldsymbol{\phi}_i + \boldsymbol{\xi}_i) + \boldsymbol{\beta}_i), \nabla \times (I - P_{h,\omega_i}^1)(\phi - P_H^1\phi)) \tag{20}$$

$$- \sum_{i \in \mathcal{N}} (\frac{1}{a}(\Sigma_c^i(\boldsymbol{\phi}_i + \boldsymbol{\xi}_i) + \boldsymbol{\beta}_i), \nabla \times (P_{h,\omega_i}^1 - P_{h,\omega_i,0}^1)(\phi - P_H^1\phi)),$$

where I is the identity operator. At this point we use Green's formula for the curl operator, that a is piecewise constant on the mesh and $\partial v_2/\partial x - \partial v_1/\partial y = 0$ for all $\boldsymbol{v} = [v_1, v_2] \in \mathscr{RT}_h(\omega)$ (since v_1 independent of y and v_2 independent of x for first order Raviart Thomas functions), and equation (10) to get,

$$\mathrm{II} \leq C \sum_{i \in \mathcal{N}} \sum_{K \in \mathcal{K}_h(\omega_i)} \| [\boldsymbol{t} \cdot \frac{1}{a}(\Sigma_c^i(\boldsymbol{\phi}_i + \boldsymbol{\xi}_i) + \boldsymbol{\beta}_i)] \|_{L^2(\partial K)} \tag{21}$$

$$\cdot \| (I - P_{h,\omega_i}^1)(\phi - P_H^1\phi) \|_{L^2(\partial K)}$$

$$+ C \sum_{i \in \mathcal{N}} \| \tilde{\Sigma}^i \|_{L^2(\partial\omega_i)} \| P_{h,\omega_i}^1(\phi - P_H^1\phi) \|_{L^2(\partial\omega_i)}.$$

We use a trace inequality, that P^1 is stable in H^1, and the Scott-Zhang interpolation estimate (see [7]), to obtain

$$\begin{aligned}
\mathrm{II} \leq & C \sum_{i \in \mathcal{N}} \sum_{K \in \mathcal{K}_h(\omega_i)} h_K \|[\boldsymbol{t} \cdot \frac{1}{a}(\Sigma_c^i(\boldsymbol{\phi}_i + \boldsymbol{\xi}_i) + \boldsymbol{\beta}_i)]\|_{L^2(\partial K)} \|\nabla \phi\|_{L^2(K)} \qquad (22) \\
& + C \sum_{i \in \mathcal{N}} H^{1/2} \|\tilde{\Sigma}^i\|_{L^2(\partial \omega_i)} \|\nabla \phi\|_{L^2(\omega_i)} \\
\leq & C \left(\sum_{i \in \mathcal{N}} \sum_{K \in \mathcal{K}_h(\omega_i)} h_K \|[\boldsymbol{t} \cdot \frac{1}{a}(\Sigma_c^i(\boldsymbol{\phi}_i + \boldsymbol{\xi}_i) + \boldsymbol{\beta}_i)]\|^2_{L^2(\partial K)} \right)^{1/2} \qquad (23) \\
& \cdot \| \frac{1}{\sqrt{a}}(\boldsymbol{\sigma} - \boldsymbol{\Sigma})\|_{L^2(\Omega)} \\
& + C \left(\sum_{i \in \mathcal{N}} H \|\tilde{\Sigma}^i\|^2_{L^2(\partial \omega_i)} \right)^{1/2} \| \frac{1}{\sqrt{a}}(\boldsymbol{\sigma} - \boldsymbol{\Sigma})\|_{L^2(\Omega)}.
\end{aligned}$$

The theorem follows immediately.

We can now present an adaptive algorithm (Algorithm 3).

Algorithm 3

1: Start with the original mesh partition with 1-layer patches and zero level of refinements everywhere.
2: Compute $\boldsymbol{\Sigma}$.
3: Compute the terms in the right hand side of (9) and set

$$\begin{aligned}
\eta_{r,i} = & \sum_{K \in \mathcal{K}_h(\omega_i)} h_K^2 \|f \psi_i + \nabla \cdot (\Sigma_c^i(\boldsymbol{\phi}_i + \boldsymbol{\xi}_i) + \boldsymbol{\beta}_i)\|^2_{L^2(K)} \qquad (24a) \\
& + \sum_{K \in \mathcal{K}_h(\omega_i)} h_K \|[\boldsymbol{t} \cdot \frac{1}{a}(\Sigma_c^i(\boldsymbol{\phi}_i + \boldsymbol{\xi}_i) + \boldsymbol{\beta}_i)]\|^2_{L^2(\partial K)} \\
\eta_{l,i} = & H \|\tilde{\Sigma}^i\|^2_{L^2(\partial \omega_i)} \qquad (24b)
\end{aligned}$$

4: Modify the mesh according to the following:

- Increase the level of refinements on patch i if $\eta_{r,i} \geq \alpha_r max_i \eta_{r,i}$.
- Increase the number of layers on patch i if $\eta_{l,i} \geq \alpha_l max_i \eta_{l,i}$.

5: Stop when $\| \frac{1}{\sqrt{a}}(\boldsymbol{\sigma} - \boldsymbol{\Sigma})\|^2_{L^2(\Omega)} / \| \frac{1}{\sqrt{a}} \boldsymbol{\sigma}\|^2_{L^2(\Omega)} \leq$ TOL, where TOL is a given tolerance.

5 Numerical Example

In the numerical example we consider the domain $\Omega = [0, 1] \times [0, 1]$ and as our coarse mesh we use a Delaunay triangulation of Ω with $H \leq 0.1$. The fine scale mesh is obtained after two regular refinements of the coarse mesh.

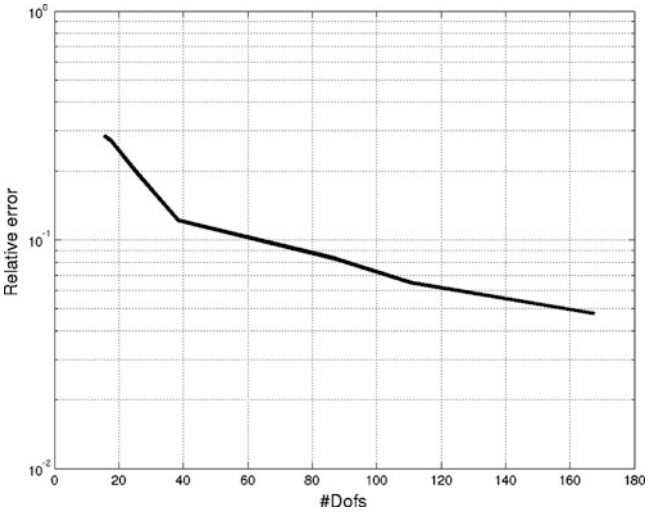


Fig. 1 The relative error approaching the desired tolerance in the adaptive algorithm

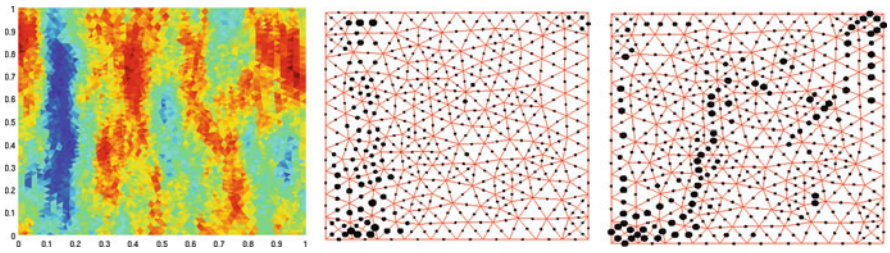


Fig. 2 The permeability function in log-scale and the adapted mesh. The *thickness* of the discs indicate the number of layers (in the *middle*) and number of refinements (to the *right*)

We let $f = 1$ on the two fine scale elements in the lower left corner, $f = -1$ on the two fine scale elements in the upper right corner, and $f = 0$ everywhere else. Thus our problem can be seen as modeling an oil reservoir with injector at the lower left corner and producer at the upper right corner of the domain.

For the permeability a, we use a permeability from the tenth comparative SPE project. See Fig. 2 for illustration of the permeability we use.

We use the adaptive Algorithm 3 and study the convergence of $\boldsymbol{\Sigma}$. As a reference solution for $\boldsymbol{\sigma}$ we use a multiscale solution computed on a mesh with three level of refinements and four layers on all patches. We use $\alpha_r = \alpha_l = 0.1$ and TOL $= 0.05$. The convergence plot is found in Fig. 1, with relative error plotted against average number of degrees of freedoms in the patches. We see that in average about 160 degrees of freedoms in the patches are required to reach the desired tolerance. A uniformly refined mesh with two levels of refinement and two

layers for all patches only yielded a relative error of 0.068 with 371 degrees of freedom in average, i.e the adaptive procedure yields a far more efficient mesh than a uniformly refined mesh. Considering that relatively few iterations are needed to reach the desired tolerance the adaptive procedure is also time efficient. The resulting adaptively refined mesh is found in Fig. 2. It is apparent and somewhat expected that it is important to refine close to the inlet and outlet. It is also clearly seen that the refinement follows the flow, i.e where the permeability is large.

References

1. Y. R. Efendiev, T. Y. Hou, and X. H. Wu, *Convergence of a nonconforming multiscale finite element method*, SIAM J. Num. Anal., 37, (2000), 888–910.
2. T. J. R. Hughes, *Multiscale phenomena: Green's functions, the Dirichlet-to-Neumann formulation, subgrid scale models, bubbles and the origins of stabilized methods,* Comput. Methods Appl. Mech. Engrg. 127, (1995), 387–401.
3. M. G. Larson and A. Målqvist, *Adaptive variational multiscale methods based on a posteriori error estimation: Energy norm estimates for elliptic problems,* Comput. Methods Appl. Mech. Engrg. 196, (2007), 2313–2324.
4. M. G. Larson and A. Målqvist, *A mixed adaptive variational multiscale method with applications in oil reservoir simulation,* Math. Models Methods Appl. Sci. 19, (2009), 1017–1042.
5. M. G. Larson and A. Målqvist, *Adaptive variational multiscale method of convection-diffusion problems*, Comm. Num. Methods Engrg. 25, (2009), 65–79,
6. D. Boffi, F. Brezzi, L. F. Demkowicz, R. G. Durán, R. S. Falk and M. Fortin, *Mixed finite elements, compatibility conditions, and applications* Springer Verlag, (2006)
7. L. R. Scott and S. Zhang, *Finite element interpolation of nonsmooth functions satisfying boundary conditions* Math. Comp, Vol. 54, No. 190, (1990), 483–493.

Adaptive Geometrical Multiscale Modeling for Hydrodynamic Problems

L. Mauri, S. Perotto, and A. Veneziani

Abstract Hydrodynamic problems often feature geometrical configurations that allow a suitable dimensional model reduction. One-dimensional models may be sometimes accurate enough for describing a dynamic of interest. In other cases, localized relevant phenomena require more precise models. To improve the computational efficiency, geometrical multiscale models have been proposed, where reduced (1D) and complete (2D–3D) models are coupled in a unique numerical solver. In this paper we consider an adaptive geometrical multiscale modeling: the regions of the computational domain requiring more or less accurate models are automatically and dynamically selected via a heuristic criterion. To the best of our knowledge, this is a first example of automatic geometrical multiscale model reduction.

1 Introduction

Typically, in hydrodynamic problems high activity regions, characterized by a wide range of spatial scales (due to shocks, wave fronts, etc.), alternate with zones where the dynamics occurs mostly along the mainstream. Due to this heterogeneity of

L. Mauri (✉)
Arianet s.r.l., Via Gilino 9, I-20128 Milano, Italy
e-mail: lorenzomauri@yahoo.it

S. Perotto
MOX, Dipartimento di Matematica "F. Brioschi", Politecnico di Milano, Piazza Leonardo da Vinci 32, I-20133 Milano, Italy
e-mail: simona.perotto@polimi.it

A. Veneziani
Department of Mathematics and Computer Science, Emory University, 400 Dowman Dr., 30322, Atlanta, GA, USA
e-mail: ale@mathcs.emory.edu

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_76,

dynamics, a high-dimensional (complete) model is often strictly required in a small portion of the computational domain whereas a low-dimensional (reduced) model is usually sufficient elsewhere. In this paper the reference hydrodynamic model is represented by the classical shallow water equations (SWE), used to describe several physical problems of interest in environmental and hydraulic engineering (e.g., tidal flows, open channel flows, free surface flows caused by dam breaking).

The simultaneous presence of heterogeneous dynamics prompted us to resort to the so-called *geometrical multiscale* reduction (see [7]), where dimensionally heterogeneous models are coupled in order to reduce the computational costs of the simulation without affecting the overall accuracy. A similar approach has been advocated in other engineering fields, like gas dynamics in internal combustion engines and computational hemodynamics (see, e.g., [2], Chap. 11). As shown in [7], the selection of the areas associated with the different models is often a challenging task, especially in the presence of fast transients. This choice is usually done a priori, driven by physical considerations. The main limitation is that a non optimal assignment of the 2D and 1D areas may either affect the accuracy of the computation when 1D equations are solved in regions where a complete model would be necessary; or affect the efficiency of the computation when the solution of the complete model is actually redundant.

Aim of this paper is to provide a criterion for an automatic selection of the 2D and 1D areas. We define a heuristic modeling error indicator based on the flow fluctuations across the control volume boundaries. Then, driven by this indicator, we set a model adaptive procedure. Preliminary results suggest that the automatic procedure improves the efficiency of the geometrical multiscale model reduction in comparison with the *a priori* splitting.

2 The Shallow Water Equations

SWE are obtained by integrating the Reynolds-averaged Navier-Stokes equations over the depth of the fluid and by assuming hydrostatic pressure distribution [10]. They express the conservation of mass and momentum for an incompressible fluid with a free surface. Getting rid of viscosity, turbulence effects and the Coriolis force, the conservative form of SWE reads

$$\frac{\partial \mathbf{u}}{\partial t} + \nabla \cdot \mathbf{F} = \mathbf{s} \qquad \text{in } \Omega, \tag{1}$$

where $\mathbf{u} = \mathbf{u}(x, y, t)$ is the vector of the conserved variables, $\mathbf{F} = \mathbf{F}(\mathbf{u})$ is the convective flux and $\mathbf{s}$ is the source term. In 2D these quantities are defined by

$$\mathbf{u} = \begin{bmatrix} h \\ hv \\ hw \end{bmatrix}, \quad \mathbf{F} = \begin{bmatrix} hv & hw \\ hv^2 + \frac{1}{2}gh^2 & hvw \\ hvw & hw^2 + \frac{1}{2}gh^2 \end{bmatrix}, \quad \mathbf{s} = \begin{bmatrix} 0 \\ -gh\left(\frac{\partial b}{\partial x} + \frac{m^2 v\sqrt{v^2+w^2}}{r^{4/3}}\right) \\ -gh\left(\frac{\partial b}{\partial y} + \frac{m^2 w\sqrt{v^2+w^2}}{r^{4/3}}\right) \end{bmatrix},$$

where h is the water depth, v and w are the (horizontal) depth-averaged velocity components along the x- and y- direction, respectively, g is the acceleration due to the gravity, b measures the bottom elevation with respect to a fixed reference level, m is the Manning coefficient due to the bed roughness and r is the hydraulic radius.

In a one-dimensional setting, equation (1) still holds provided that the definitions of $\mathbf{u}$, $\mathbf{F}$ and $\mathbf{s}$ simplify in

$$\mathbf{u} = \begin{bmatrix} h \\ hv \end{bmatrix}, \quad \mathbf{F} = \begin{bmatrix} hv \\ hv^2 + \frac{1}{2}gh^2 \end{bmatrix}, \quad \mathbf{s} = \begin{bmatrix} 0 \\ -gh\left(\frac{db}{dx} + \frac{m^2 v^2}{r^{4/3}}\right) \end{bmatrix},$$

respectively. Appropriate initial and boundary conditions depending on the considered hydrodynamic configuration complete equations (1).

In view of the numerical simulation of system (1), Godunov-type finite volume schemes are largely employed on both structured and unstructured grids (see, e.g., [3, 4, 9]). Here we use a structured quadrilateral grid $\mathscr{T}$, with a second-order Godunov-type scheme based on the Roe linearized Riemann solver and the super bee flux limiters (see, e.g., [4, 8]), combined with the 2D *corner transport upwind* (CTU) method due to Colella for multidimensional integration [1].

Let the cells of $\mathscr{T}$ be identified by the pairs (i, j), being $i(j)$ the cell index in the $x(y)$-direction, with the notation $\mathscr{C}_{i,j} = [x_{i-1/2}, x_{i+1/2}] \times [y_{j-1/2}, y_{j+1/2}]$. Index n refers to the time level; Δx, Δy and Δt denote the uniform size of $\mathscr{T}$ along the x- and y-direction and the time step; $\mathbf{U}_{i,j}^n$ is the numerical approximation to $\frac{1}{\Delta x\, \Delta y} \int_{\mathscr{C}_{i,j}} \mathbf{u}(x, y, t^n)\, dxdy$. The finite volume discretization in the wave propagation form reads

$$\begin{aligned} \mathbf{U}_{i,j}^{n+1} = \mathbf{U}_{i,j}^n &- \frac{\Delta t}{\Delta x}\left(\mathscr{A}^+ \Delta \mathbf{U}_{i-1/2,j}^n + \mathscr{A}^- \Delta \mathbf{U}_{i+1/2,j}^n\right) \\ &- \frac{\Delta t}{\Delta y}\left(\mathscr{B}^+ \Delta \mathbf{U}_{i,j-1/2}^n + \mathscr{B}^- \Delta \mathbf{U}_{i,j+1/2}^n\right) \\ &- \frac{\Delta t}{\Delta x}\left(\mathscr{F}_{i+1/2,j}^n - \mathscr{F}_{i-1/2,j}^n\right) - \frac{\Delta t}{\Delta y}\left(\mathscr{G}_{i,j+1/2}^n - \mathscr{G}_{i,j-1/2}^n\right) + \Delta t\, \mathbf{S}_{i,j}^n. \end{aligned} \tag{2}$$

Since scheme (2) is explicit, we select a time step Δt to fulfill the CFL condition.

The horizontal fluctuations

$$\mathscr{A}^{\pm} \Delta \mathbf{U}_{i\mp 1/2,j}^n = \sum_{p=1}^{3} \left(s_{i\mp 1/2,j}^p\right)^{\pm} \mathscr{W}_{i\mp 1/2,j}^p \tag{3}$$

measure the net effect of all the right-going waves $\mathscr{W}_{i-1/2,j}^p$ from the interface $\{x = x_{i-1/2}\} \times [y_{j-1/2}, y_{j+1/2}]$ with speed $s_{i-1/2,j}^p$ and of all the left-going waves $\mathscr{W}_{i+1/2,j}^p$ from the interface $\{x = x_{i+1/2}\} \times [y_{j-1/2}, y_{j+1/2}]$ with speed $s_{i+1/2,j}^p$, respectively; likewise, the vertical fluctuations

$$\mathscr{B}^{\pm}\Delta\mathbf{U}^{n}_{i,j\mp1/2}=\sum_{p=1}^{3}\left(s^{p}_{i,j\mp1/2}\right)^{\pm}\mathscr{W}^{p}_{i,j\mp1/2} \tag{4}$$

take into account the net effect of all the up-going waves $\mathscr{W}^{p}_{i,j-1/2}$ from the interface $[x_{i-1/2},x_{i+1/2}]\times\{y=y_{j-1/2}\}$ with speed $s^{p}_{i,j-1/2}$ and of all the down-going waves $\mathscr{W}^{p}_{i,j+1/2}$ from the interface $[x_{i-1/2},x_{i+1/2}]\times\{y=y_{j+1/2}\}$ with speed $s^{p}_{i,j+1/2}$, respectively. Notice that all the fluctuations in (3) and (4) are suitably modified via the Harten-Hyman entropy fix correction to treat also transonic rarefaction waves [4].

Terms $\mathscr{F}^{n}_{i\pm1/2,j}$ and $\mathscr{G}^{n}_{i,j\pm1/2}$ in (2) include the effects related to the waves transversely propagating from the neighboring cells into $\mathscr{C}_{i,j}$: they can be distinguished into the left-going and right-going transverse waves

$$\mathscr{F}^{n}_{i\pm1/2,j}=-\frac{\Delta t}{2\Delta y}\mathscr{A}^{\pm}\Big[\mathscr{B}^{-}\Delta\mathbf{U}^{n}_{i,j+1/2}+\mathscr{B}^{+}\Delta\mathbf{U}^{n}_{i,j-1/2}\Big],$$

respectively, and into the down-going and up-going transverse waves

$$\mathscr{G}^{n}_{i,j\pm1/2}=-\frac{\Delta t}{2\Delta x}\mathscr{B}^{\pm}\Big[\mathscr{A}^{-}\Delta\mathbf{U}^{n}_{i+1/2,j}+\mathscr{A}^{+}\Delta\mathbf{U}^{n}_{i-1/2,j}\Big],$$

respectively (see Fig. 1). These corrections are first-order accurate. Hereafter we resort to a second order extension (see [4], Chap. 20).

The source term is integrated via a fractional step method (the *Godunov splitting*). We first solve the SWE with no source term on the time interval $I_n=[t^n,t^{n+1})$, with initial datum $\mathbf{U}^n=[\mathbf{U}^n_{i,j}]$; this predictor step yields an intermediate solution $\mathbf{U}^{n+1,*}$. Then, we solve the independent system of ODEs $\partial\mathbf{u}/\partial t=\mathbf{s}$ on each cell and on I_n, with initial datum $\mathbf{U}^{n+1,*}$. This corrector step provides the approximation $\mathbf{U}^{n+1}$. In particular, we use an explicit second-order Runge-Kutta scheme to solve the system of ODEs: fulfillment of the CFL condition is guaranteed by an appropriate selection of the time step.

Two types of boundary conditions are used herein: *nonreflecting* boundary conditions in correspondence with the open boundaries and *slip* conditions along the solid walls. In both the cases, we resort to *ghost cells*. At the beginning of each time step, the values of the solution in the ghost cells are determined by an appropriate extrapolation of the solution at the previous steps or of the boundary conditions. In particular, we add two ghost cells along $\partial\Omega$ and we employ a zero-order extrapolation to set values here (see [4] for further details).

3 The Adaptive Geometrical Multiscale Solver

In this section we consider a geometrical multiscale formulation. This means that equations (1) for both the 1D and 2D domains are numerically coupled. Our goal is an *automatic* detection of the areas of Ω where the water dynamics needs to be

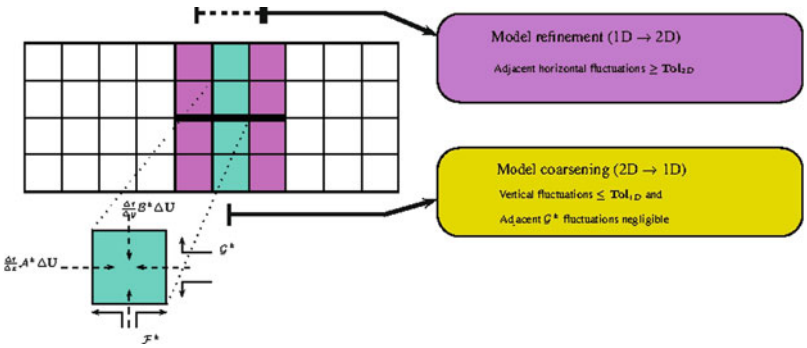


Fig. 1 Fluctuations driving the modeling adaptive procedure

described with a 2D model or can be modelled by a 1D problem. To this aim, we need a modeling error indicator; then we set a modeling adaptive procedure driven by such an indicator to get the 2D–1D SW model.

3.1 Heuristic Modeling Error Indicator for the SWE

Two different error indicators are devised to estimate: (a) in 2D subdomains the possible reduction to 1D (model coarsening); (b) in 1D subdomains the possible expansion to 2D (model refinement). They are both heuristic indicators and are based on the concept of fluctuation introduced in (3)–(4). In particular, we assume as reference hydrodynamic configuration an open rectilinear channel characterized by a constant rectangular cross-section and subdivided into $N_x \times N_y$ cells, slip conditions being assigned along the solid walls.

Model coarsening. Let us focus on the i-th column of cells and, in particular, on relation (2) which updates the value of $\mathbf{U}$ on the cell $\mathscr{C}_{i,j}$ at $t = t^{n+1}$. For simplicity, we assume here that the source term $\mathbf{S}^n_{i,j}$ is zero.

As indicator driving the model coarsening we assume the value $|\mathscr{B}^{\pm}\Delta\mathbf{U}^n_{i,j\mp1/2}|$ of the vertical fluctuations: if they are small, i.e.,

$$|\mathscr{B}^{\pm}\Delta\mathbf{U}^n_{i,j\mp1/2}| \leq \mathbf{Tol}_{1D} \qquad \forall\, j = 1,\ldots,N_y, \tag{5}$$

then the i-th column is marked to be assigned to the 1D model at $t = t^{n+1}$ (see Fig. 1). Notice that $\mathbf{Tol}_{1D}$ is in general a 3-component vector, since the coarsening check has to be tested for each component of $\Delta\mathbf{U}^n_{i,j\mp1/2}$.

Let us provide some rationale behind criterion (5), referring to [6] for further details. Let us assume that the fluctuations $\mathscr{B}^{\pm}\Delta\mathbf{U}^n_{i,j\mp1/2}$ are identically equal to zero, for $j = 1,\ldots,N_y$. Then, the second line of (2) and the terms $\mathscr{F}^n_{i\pm1/2,j}$ vanish. To decide whether also the down-going and up-going transverse waves may be ignored, we consider the $(i-1)$-th and $(i+1)$-th columns, where we assume that

the vertical fluctuations are equal to zero as well. It can be empirically inferred that, if $\mathscr{B}^{\pm}\Delta\mathbf{U}^n_{k,j\mp1/2}=0$ for $k=i-1,i,i+1$ and $j=1,\ldots,N_y$, then, in those three columns w vanishes while the water depth h is column-wise constant. Physically, this is consistent with two cases: (*i*) h and v are constant along the x-direction: the horizontal fluctuations $\mathscr{A}^{\pm}\Delta\mathbf{U}^n_{i\mp1/2,j}$ and the transverse waves $\mathscr{G}^n_{i,j\pm1/2}$ vanish, so that equation (2) reduces to $\mathbf{U}^{n+1}_{i,j}=\mathbf{U}^n_{i,j}$ and no vertical fluctuation is expected in the i-th column at time $t=t^{n+1}$; (*ii*) h and v are constant along the y-direction: the horizontal fluctuations $\mathscr{A}^{+}\Delta\mathbf{U}^n_{i-1/2,j}$ and $\mathscr{A}^{-}\Delta\mathbf{U}^n_{i+1/2,j}$ have constant values with respect to j, so that $\mathscr{G}^n_{i,j+1/2}=\mathscr{G}^n_{i,j-1/2}=0$; equation (2) thus reduces to

$$\mathbf{U}^{n+1}_{i,j}=\mathbf{U}^n_{i,j}-\frac{\Delta t}{\Delta x}\Big(\mathscr{A}^{+}\Delta\mathbf{U}^n_{i-1/2,j}+\mathscr{A}^{-}\Delta\mathbf{U}^n_{i+1/2,j}\Big), \tag{6}$$

and no vertical fluctuation is expected in the i-th column at time $t=t^{n+1}$. Should h and v vary along both x- and y-direction, the four terms $\mathscr{A}^{\pm}\Delta\mathbf{U}^n_{i\mp1/2,j}$, $\mathscr{G}^n_{i,j\pm1/2}$ do not vanish and vary for different values of j. So we infer that this assumption is not compatible with the condition that the vertical fluctuations in the current and adjacent columns vanish.

In practice, condition $\mathscr{B}^{\pm}\Delta\mathbf{U}^n_{i,j\mp1/2}=0$ is relaxed and turns into criterion (5). If this check holds for the $(i-1)$-th, i-th and $(i+1)$-th column, then the i-th column is eligible to be associated with the 1D model. At this point equation (6) is solved by assuming for $\mathbf{U}^n_{k,j}$, with $k=i,i\pm1/2$, the mean value of $\mathbf{U}^n_{k,j}$ over the k-th column.

Model refinement. It is less immediate to find a reliable heuristic criterion driving the refinement of the 1D areas. In principle one should quantify the vertical fluctuations but, of course, these quantities are not computed by a one-dimensional model. Nevertheless, high horizontal fluctuations in the 2D domains adjacent to a 1D area are likely to transfer energy along the vertical direction, triggering a significant vertical component of the velocity. This is the reason why we focus on the fluctuations in the 2D columns neighboring with a 1D segment. In more detail, if the i-th column is adjacent on the right to a 1D segment and $|\mathscr{A}^{+}\Delta\mathbf{U}^n_{i+1/2,j}|$ is sufficiently large, for some $j=1,\ldots,N_y$, we expect to have a significant vertical component in the neighbors, as would be triggered by the $\mathscr{G}-$terms in (2) for a full 2D setting. The corresponding 1D segment becomes consequently a candidate to be a (the $(i+1)$-th) 2D column; likewise if the i-th column is adjacent, on the left, to a 1D segment and $|\mathscr{A}^{-}\Delta\mathbf{U}^n_{i-1/2,j}|$ is sufficiently large, for some $j=1,\ldots,N_y$, then the 1D segment becomes eligible to be a (the $(i-1)$-th) 2D column (see Fig. 1).

The error indicator driving the model refinement is consequently represented by the horizontal fluctuations $\mathscr{A}^{\pm}\Delta\mathbf{U}^n_{i\pm1/2,j}$; the corresponding refinement criterion reads: if

$$|\mathscr{A}^{\pm}\Delta\mathbf{U}^n_{i\pm1/2,j}|\geq\mathbf{Tol}_{2D}\qquad\text{for some } j=1,\ldots,N_y, \tag{7}$$

then the 1D segment at the right (at the left) of the i-th column is marked to be assigned to the 2D model at $t=t^{n+1}$. As in (5), $\mathbf{Tol}_{2D}\in\mathbb{R}^3$ and the refinement check has to be verified for each component of $\Delta\mathbf{U}^n_{i\pm1/2,j}$.

3.2 The Modeling Adaptive Procedure

This procedure moves from the coarsening and the refinement criteria above to set up an automatic selection of the 2D and 1D areas. We can itemize the generic k-step of the modeling adaptive procedure we propose in such a way: running over all the N_x columns,

1. We mark all the 2D columns where criterion (5) holds as eligible for the 1D model and we group the consecutive columns thus marked;
2. We mark all the 1D intervals neighboring with a 2D column where one of the criteria (7) holds as eligible for the 2D model and we group the consecutive intervals thus marked;
3. We select the groups in 1. and 2. which are neighbors with sets of at least $min2d$ 2D columns and/or sets of at least $min1d$ subintervals;
4. All the groups identified by 1. and 3. constituted by at least q columns, with $q \geq min1d$, are assigned to the 1D model;
5. All the groups identified by 2. and 3. constituted by at least p subintervals, with $p \geq min2d$, are assigned to the 2D model.

This approach guarantees always a minimum size for both the 2D and 1D areas given by $min2d \cdot \Delta y$ and $min1d \cdot \Delta x$, respectively. Moreover, we permanently associate a certain area of the domain with the 2D model in the presence of a hydrodynamic configuration (e.g., a pillar, a pier) or a boundary condition (e.g., a lateral inlet) which implicitly induces vertical fluctuations.

Concerning the matching conditions between the two classes of models, we distinguish *1D-2D* and *2D-1D couplings*. For the former, we extend the 1D values of h and hv to all the N_y cells in the first column of the 2D domain, while setting $hw = 0$. For the latter, the mean value of h and hv over the N_y cells in the last column of the 2D domain is assigned to the corresponding 1D variables. Notice that empirical criteria (5) and (7) rely on the assumption of no-forcing term. Should a forcing term be on, these procedures need to be properly modified [6].

The same time step Δt for both the 2D and the 1D domains is selected so that the CFL condition is globally fulfilled. Finally, to contain the computational cost of the whole adaptive procedure, we update the 2D/1D model every M^* time-steps instead at each time step.

4 A Numerical Example

With this test case we analyze the proposed modeling adaptive procedure, essentially from a qualitative viewpoint. We have implemented the adaptive solver in Clawpack 4.3 [5]. We consider a popular hydrodynamic benchmark, i.e., a rectangular dam-break symmetrically localized in a 10m $\times$3m rectangular channel with a flat horizontal frictionless bed (see Fig. 2, top-left). The dam break occurs due to the istantaneous collapse of three of the dam walls. We employ a grid consisting of

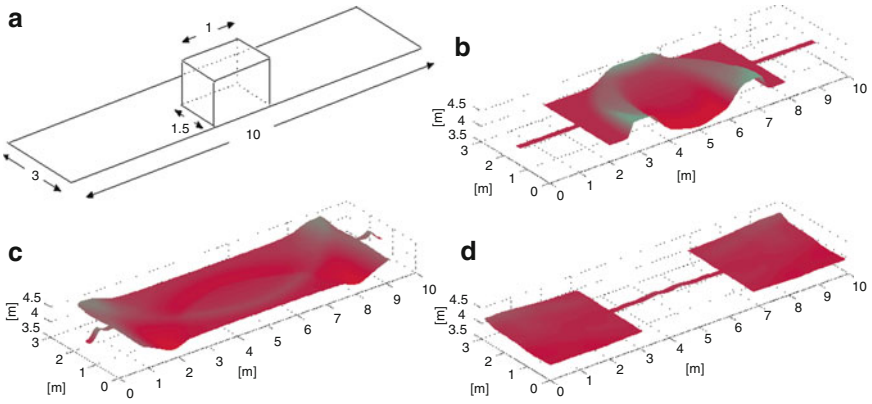


Fig. 2 Adaptive geometrical multiscale modeling: test case sketch (*top-left*); 3D representation of the free surface at $t = 1$ s (*top-right*), $t = 2$ s (*bottom-left*), $t = 10$ s (*bottom-right*)

100×30 cells, we assign slip boundary conditions on the whole $\partial\Omega$, and we set 1 as maximum value for the CFL condition. Concerning the parameters involved in the modeling adaptive procedure, we set: $\mathbf{Tol}_{1D} = \mathbf{Tol}_{2D} = [10^{-1}, 10^{-1}, 10^{-1}]^T$, $min2d = 10$, $min1d = 2$, $M^* = 5$.

In Fig. 2 we show the water surface for three different times: the 2D model follows the evolution of the dynamics as well as that the intrisic symmetry of the problem is preserved by the model adaptation.

For a more quantitative investigation of the modeling adaptive procedure we refer to [6].

References

1. Colella, P.: Multidimensional upwind methods for hyperbolic conservation laws. J. Comput. Phys. **87**, 171–200 (1990)
2. *Cardiovascular Mathematics, Modeling and Simulation of the Circulatory System*. Formaggia, L., Quarteroni, A., Veneziani, A. (eds.), Springer, Milano (2009)
3. Krámer, T., Józsa, J.: Solution-adaptivity in modelling complex shallow flows. Computers & Fluids **36**, 562–577 (2007)
4. LeVeque, R.J.: *Finite Volume Methods for Hyperbolic Problems*. Cambridge, (2001).
5. LeVeque, R.J.: Clawpack, Version 4.3. http://depts.washington.edu/clawpack/clawpack-4.3/
6. Mauri, L., Perotto, S., Veneziani, A.: An adaptive geometrical multiscale model for the shallow water equations. In preparation (2012)
7. Miglio, E., Perotto, S., Saleri, F.: Model coupling techniques for free-surface flow problems. Part I. Nonlinear Analysis, **63**, 1885–1896 (2005)
8. Roe, P.L.: Approximate Riemann solvers, parameter vectors, and difference schemes. J. Comput. Phys. **43**, 357–372 (1981)
9. Sleigh, P.A., Berzins, M., Gaskell, P.H., Wright, N.G.: An unstructured finite volume algorithm for predicting flow in rivers and estuaries. Computers & Fluids **27**, 479–508 (1998)
10. Vreugdenhil, C.B.:*Numerical Methods for Shallow-Water Flow*. Springer, (1994).

On the Superlinear Convergence of MINRES

V. Simoncini and D.B. Szyld

Abstract Quantitative bounds are presented for the superlinear convergence of the MINRES method of Paige and Saunders (SIAM J Numer Anal 12:617–629, 1975) for the solution of sparse linear systems $Ax = b$, with A symmetric and indefinite. It is shown that the superlinear convergence is observed as soon as the harmonic Ritz values approximate well the eigenvalues of A that are either closest to zero or farthest from zero. This generalizes a well-known corresponding result obtained by van der Sluis and van der Vorst with respect to the Conjugate Gradients method, for A symmetric and positive definite.

1 Introduction

The MINRES method is a short-term recurrence Krylov subspace method developed by Paige and Saunders [8] for the solution of large and sparse linear systems of equations of the form

$$Ax = b, \tag{1}$$

where the $n \times n$ matrix A is symmetric and indefinite. MINRES is in fact very popular for solving *indefinite* linear systems, and it has become the leading solver for symmetric saddle point linear systems, for which spectral information can often

V. Simoncini (✉)
Dipartimento di Matematica, Università di Bologna, Piazza di Porta S. Donato, 5, I-40127, Bologna, Italy

CIRSA, Ravenna, Italy
e-mail: valeria.simoncini@unibo.it

D.B. Szyld
Department of Mathematics, Temple University (038-16), 1805 N. Broad Street, Philadelphia, PA, 19122-6094, USA
e-mail: szyld@temple.edu

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_77,

be obtained from the application problem; see, e.g., [2] and references therein. It is well-known that MINRES exhibits superlinear convergence, i.e., that the norm of the residuals decreases linearly at first, but then, as the iterations progress the linear rate accelerates (cf. Fig. 2). The motivation of this paper is to explain this observed phenomenon. We show that the superlinear convergence behavior (i.e., the change of the linear rate) occurs when the harmonic Ritz values approximate well the eigenvalues of the matrix A that are closest to or farthest away from the origin. This is consistent with the exposition in [7, Sect. 7], and with the comments found in [16, p. 78]. We are interested in describing a quantitative bound explaining more precisely these observations.

After a brief description of the algorithm in Sect. 2, we collect different results on the convergence of MINRES available in the literature (Sect. 3) and, inspired by ideas from other contexts, we develop a quantitative bound for its superlinear convergence (Sect. 4).

Throughout the paper exact arithmetic is assumed.

2 Review and Preliminaries

We review here some concepts which we use throughout the paper. Given a first approximation x_0 to the solution of (1), and the corresponding initial residual $r_0 = b - Ax_0$, the Krylov subspace of dimension m defined by A and r_0 is given by

$$\mathscr{K}_m = \mathscr{K}_m(A, r_0) = \text{span}\{r_0, Ar_0, A^2r_0, \ldots, A^{m-1}r_0\}. \tag{2}$$

An orthonormal basis $\{v_1, \ldots, v_m\}$ of $\mathscr{K}_m$ can be built by means of the Lanczos method. Let $V_m = [v_1, \ldots, v_m]$ collect these vectors, and observe that the matrix $T_m = V_m^T A V_m$ is symmetric and tridiagonal; the latter property is a consequence of the three-term recurrence from the Lanczos process. For details, see, e.g., [4,8,9,13].

Like many other projection-type approaches, at the mth step an approximation to the solution of (1), x_m, can be obtained in $x_0 + \mathscr{K}_m$, by imposing some additional condition. In MINRES, this approximation is found by requiring that the norm of the corresponding residual $r_m = b - Ax_m$ is minimized over all possible vectors of the form $x_m = x_0 + z$, with $z \in \mathscr{K}_m$; here and in the following we shall only consider the Euclidean norm, although the use of other norms has been analyzed in the literature; see, e.g., [10]. Thus, this approximation is of the form $x_m = x_0 + q_{m-1}(A)r_0$, where q_{m-1} is a polynomial of degree at most $m-1$. This implies that the residual $r_m = b - Ax_m$ is associated with the so-called *residual polynomial* $p_m(t)$ of degree at most m with $p_m(0) = 1$, since $r_m = b - Ax_m = r_0 - Aq_{m-1}(A)r_0 = p_m(A)r_0$. We recall two sets of scalars approximating the eigenvalues of the matrix A, as the iteration progresses: The *Ritz values* (with respect to $\mathscr{K}_m$), which are the eigenvalues of T_m, and the *harmonic Ritz values*, which instead are the roots of the residual polynomial $p_m(t)$, and are denoted by $\theta_1^{(m)}, \ldots, \theta_m^{(m)}$, i.e.,

$$p_m(t) = \frac{(\theta_1^{(m)} - t)\cdots(\theta_m^{(m)} - t)}{\theta_1^{(m)}\cdots\theta_m^{(m)}}.$$

The harmonic Ritz values can be equivalently characterized as the Ritz values of A^{-1} with respect to $A\mathscr{K}_m$; see, e.g., [5]. From a computational view point, the harmonic Ritz values can be obtained as the eigenvalues of the pencil $(\underline{T_m}^T \underline{T_m}, T_m)$, where $\underline{T_m} = V_{m+1}^T A V_m$; see [7] and references therein. As a special feature, we also notice that harmonic Ritz values approximate the eigenvalues from the interior of the spectral intervals of A. Therefore, any interval around the origin that is free of eigenvalues of A is also free of harmonic Ritz values [7]. This ensures that the approximation, say, to the smallest positive eigenvalues is genuine, and it is not incidental, since no harmonic Ritz value will cross the origin to approximate the negative eigenvalues as m increases, the way Ritz values would do, on indefinite matrices. We also mention that harmonic Ritz values may play an important role in practical circumstances, such as the approximation of interior eigenvalues, see, e.g., [5], and for devising problem-dependent stopping criteria [11].

3 Known Bounds for the Residual Norm

Let $\Lambda(A) = \{\lambda_1, \ldots, \lambda_n\}$ be the set of eigenvalues of A, with the eigenvalues ordered increasingly, and let $\mathscr{P}_m$ be the set of all polynomials p of degree at most m such that $p(0) = 1$.

From $r_m = p_m(A)r_0$, we have the following standard bound

$$\|r_m\| = \|p_m(A)r_0\| \le \min_{p\in\mathscr{P}_m} \max_{i=1,\ldots,n} |p(\lambda_i)|\,\|r_0\|. \tag{3}$$

Therefore, it is useful to find appropriate bounds for

$$E_m(\Lambda(A)) = \min_{p\in\mathscr{P}_m} \max_{\lambda\in\Lambda(A)} |p(\lambda)|,$$

and these will depend of course on the form of the set of eigenvalues $\Lambda(A)$. One such bound was developed for the case where $\Lambda(A) \subset [a, b] \cup [c, d]$, where $a < b < 0 < c < d$, under the constraint that $|b-a| = |d-c|$, that is, the two intervals have equal length. In this case, using an appropriate transformation of the intervals and bounds on Chebychev polynomials, the following bound holds:

$$\frac{\|r_m\|}{\|r_0\|} \le 2\left(\frac{\sqrt{|ad|} - \sqrt{|bc|}}{\sqrt{|ad|} + \sqrt{|bc|}}\right)^{[m/2]},$$

where $[m/2]$ is the integer part of $m/2$; see [3, Chap. 3], or [4, Sect. 3.1], for details.

Bounds for the asymptotic convergence factor $\lim_{m\to\infty} e_m^{\frac{1}{m}}$ with $e_m = E_m([a,b]\cup[c,d])$, were proposed in [18], where the role of $\sqrt{bc/ad}$ was also emphasized.

For the special case where the number of negative (or positive) eigenvalues is relatively small, say k, we can use the technique in [14, Theorem 4.4] to provide a more descriptive bound as follows.

Proposition 1. *Let $\Lambda(A) \subset \{\lambda_1,\ldots,\lambda_k\}\cup[c,d]$, with $\lambda_1,\ldots,\lambda_k$ negative, and $0 < c \le d$. Then, for $m > k$,*

$$\frac{\|r_m\|}{\|r_0\|} \le \Omega_k \frac{2}{\rho^{k-m}+\rho^{m-k}},$$

where $\rho = \dfrac{\sqrt{\tilde\kappa}+1}{\sqrt{\tilde\kappa}-1}$, $\tilde\kappa = \frac{d}{c}$, and $\Omega_k = \displaystyle\prod_{j=1}^{k}\left(1-\frac{d}{\lambda_j}\right)$ is independent of m.

Beckermann and Kuijlaars [1] developed bounds for the quantities $E_m(S)$ for very specific sets S containing the spectrum of positive definite matrices A. These bounds were useful to follow the superlinear convergence of Conjugate Gradients (CG). For description of Conjugate Gradients, or other Krylov subspace methods, see, e.g., [4, 9, 13]. Beckermann and Kuijlaars further indicated that the general results they proved would be applicable to MINRES as well, but for this one needs to build the appropriate sets S containing $\Lambda(A)$ now having negative and positive elements. Calculating these sets "is a problem in itself," and this was not developed in [1].

We note in passing that the *a posteriori* convergence bounds developed in [12] can also apply to MINRES. They are based on how close invariant subspaces of A are to the Krylov subspace; in the present context, this reduces to the angle between eigenvectors and the Krylov subspace.

4 A New A-posteriori Bound

As opposed to most a-priori estimates recalled in the previous section, here we describe a new *a-posteriori* bound that aims to describe the possibly abrupt steepness change in the linear convergence rate that is often encountered when using MINRES. Detecting and understanding this behavior may help devise an improved method, or an improved preconditioner, that allow the method to immediately enter the superlinear convergence stage without the initial slower phase; see, e.g., [6].

We show that after a sufficient number of iterations have been performed, the method behaves as if the eigencomponents corresponding to the smallest eigenvalues (in modulo) had been removed. Since the (worst case) rate of convergence depends on the spectral interval, the method behaves as if the matrix had a reduced spectral interval, hence improving its convergence rate. This

phenomenon is well known for CG, and it was completely uncovered by van der Sluis and van der Vorst in their 1986 paper [15]. We essentially take their proof for CG, which uses Ritz values, and obtain a similar result for MINRES using harmonic Ritz values. We use the same polynomial for the bound, which is also used in [17, Lemma 1.5] for the nonsymmetric case. We should also add that Van der Vorst in [16, p. 78] already mentions the possibility of developing this bound in this form. Here we present it in detail.

Let (λ_k, z_k), $k = 1, \ldots, n$ be the eigenpairs of A, with λ_k, $k = 1, \ldots, n$ sorted in increasing absolute value, and assume that λ_1 is simple.

Theorem 1. *Let r_m be the residual after m MINRES iterations with starting residual r_0, so that in particular $r_m = b - Ax_m$ with $x_m \in x_0 + \mathscr{K}_m(A, r_0)$. Let us write $r_m = \bar{r}_0 + s^{(1)}$, with $\bar{r}_0 \perp z_1$, and let $\bar{r}_j$ be the MINRES residual after j iterations in $\mathscr{K}_j(A, \bar{r}_0)$. Then after $m + j$ MINRES iterations with starting residual r_0 we obtain*

$$\|r_{m+j}\| \le F_m \|\bar{r}_j\|, \quad \text{where} \quad F_m = \max_{k \ge 2} \frac{|\theta_1^{(m)}|}{|\lambda_1|} \frac{|\lambda_1 - \lambda_k|}{|\theta_1^{(m)} - \lambda_k|}$$

and $\theta_1^{(m)}$ is the harmonic Ritz value closest to λ_1 in $\mathscr{K}_m(A, r_0)$.

Proof. Let $p_m, \bar{q}_j$ be the MINRES residual polynomials in $\mathscr{K}_m(A, r_0)$ and $\mathscr{K}_j(A, \bar{r}_0)$, respectively. We write $r_0 = \sum_{k=1}^{n} \gamma_k z_k$ so that

$$r_m = p_m(A) r_0 = \sum_{k=1}^{n} p_m(\lambda_k) \gamma_k z_k, \quad \bar{r}_0 = \sum_{k=2}^{n} p_m(\lambda_k) \gamma_k z_k.$$

Moreover, $\bar{r}_j = \bar{q}_j(A) \bar{r}_0 = \sum_{k=2}^{n} \bar{q}_j(\lambda_k) p_m(\lambda_k) \gamma_k z_k$. Let

$$\phi_m(\lambda) = \frac{\theta_1^{(m)}}{\lambda_1} \frac{\lambda_1 - \lambda}{\theta_1^{(m)} - \lambda} p_m(\lambda),$$

and notice that $\phi_m(\lambda_1) = 0$.

Since the MINRES polynomial is a minimizing polynomial, we obtain

$$\begin{aligned} \|r_{m+j}\|^2 = \|p_{m+j}(A) r_0\|^2 &\le \|\phi_m(A) \bar{q}_j(A) r_0\|^2 = \sum_{k=2}^{n} \phi_m(\lambda_k)^2 \bar{q}_j(\lambda_k)^2 \gamma_k^2 \\ &\le F_m^2 \sum_{k=2}^{n} p_m(\lambda_k)^2 \bar{q}_j(\lambda_k)^2 \gamma_k^2 = F_m^2 \|\bar{r}_j\|^2. \end{aligned}$$

□

The bound for $\|r_{m+j}\|$ shows that the residual norm can be bounded by the norm of the residual deflated of the eigenvector component corresponding to λ_1. If one of the harmonic Ritz values is a good approximation to λ_1, then the factor F_m will be very close to one. Therefore, in this case the behavior of the residual norm $\|r_{m+j}\|$ is well represented by that of $\bar{r}_j$, which has no eigencomponent onto z_1.

The result can be easily generalized to a group of eigenvalues, the only technical change would be the use of more orthogonality conditions to define $\bar{r}_0$. Nowhere in the proof we used the fact that λ_1 is the eigenvalue closest to the origin. In fact, the proof holds for any simple eigenvalue of A, and in particular for those farthest from the origin.

Example 1. We consider the following data:

$$A = \begin{bmatrix} A_- & & \\ & \delta & \\ & & A_+ \end{bmatrix}, \qquad b = \mathbf{1},$$

where $\delta = -10^{-3}$, and A_+, A_- are diagonal matrices with values logarithmically distributed in $[10^0, 10^{0.5}]$ and $[-10^1, -10^0]$, respectively. The dimension of A is $n = 2 \cdot 399 + 1 = 799$.

The convergence history of MINRES on $Ax = b$ shows a long plateau, with an almost complete stagnation (cf. Fig. 1), corresponding to the effort the method is making in approximating the interior eigenvalue δ, once it discovers there is one. This fact can be clearly observed in Fig. 2, where the values $\min_i |\theta_i^{(m)} - \delta|$ are reported, where $\theta_i^{(m)}$, $i = 1, \ldots, m$, are the harmonic Ritz values at the mth iteration.

Let r_{70} be the residual of MINRES on $Ax = b$ after 70 iterations. The dashed curve in Fig. 1 reports the convergence history of a MINRES process started with r_{70} as initial residual. Its convergence rate matches quite well that of the original MINRES after the smallest eigenvalue is singled out. For the sake of completeness, in Fig. 1 we also report the convergence history of MINRES applied to the companion problem $A_1 x_1 = b_1$ where the row and column corresponding to δ are removed. The plot shows that the convergence delay is only due to the isolated small eigenvalue.

Example 2. We next consider a spectral distribution that is possibly more common in practice, and in which the picture of superlinear convergence rate is more typical. We consider a variant of the previous example, where now $\delta = \mathrm{diag}(-10^{-1}, -3 \cdot 10^{-1}, -2 \cdot 10^{-1})$, so that the matrix A has size $n = 801$; the right-hand side is $b = \mathbf{1}$, as in Example 1. The small negative eigenvalues are now less isolated, and their approximation during the MINRES process is more effective (cf. Fig. 3). Nonetheless, as soon as the Krylov space captures the small eigenvalues – after about 70 iterations – the MINRES convergence rate changes, showing superlinear convergence.

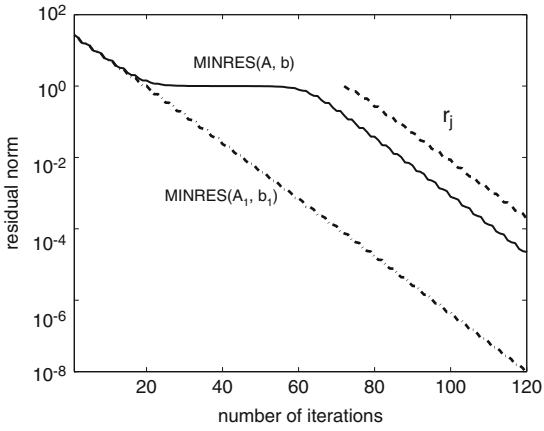


Fig. 1 Example 1. Convergence history of MINRES on $Ax = b$ and $A_1x_1 = b_1$

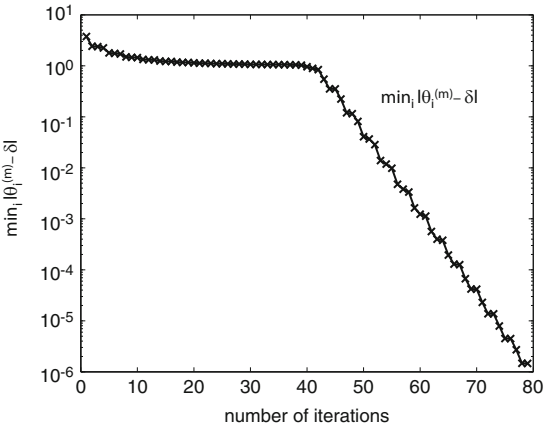


Fig. 2 Example 1. Convergence history of the harmonic Ritz value closest to δ

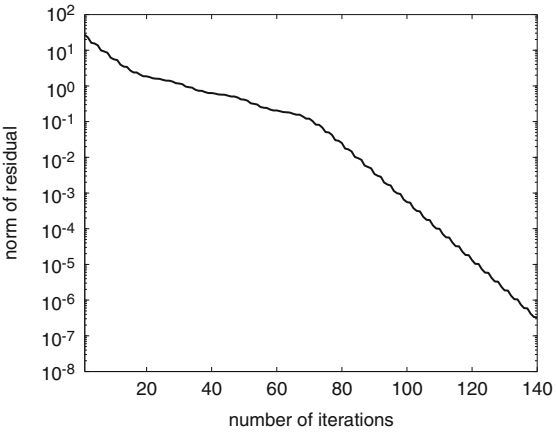


Fig. 3 Example 2. Convergence history of MINRES on $Ax = b$

Acknowledgements This research is supported in part by the U.S. National Science Foundation under grant DMS-1115520 and by the U.S. Department of Energy under grant DE-FG02-05ER25672.

References

1. Bernhard Beckermann and Arno B.J. Kuijlaars. Superlinear convergence of Conjugate Gradients. *SIAM Journal on Numerical Analysis*, 39:300–329, 2001.
2. Michele Benzi, Gene H. Golub, and Jörg Liesen. Numerical solution of saddle point problems. *Acta Numerica*, 14:1–137, 2005.
3. Bernd Fischer. *Polynomial Based Iteration Methods for Symmetric Linear Systems*. Willey and Teubner, Chichester, West Essex, England, and Stuttgart, 1996. Reprinted by SIAM, Philadelphia, 2012.
4. Anne Greenbaum. *Iterative Methods for Solving Linear Systems*, volume 17 of *Frontiers in Applied Mathematics*. SIAM, Philadelphia, 1997.
5. Ronald B. Morgan. Computing interior eigenvalues of large matrices. *Linear Algebra and its Applications*, pages 289–309, 1991.
6. Maxim A. Olshanskii and Valeria Simoncini. Acquired clustering properties and solution of certain saddle point systems. *SIAM Journal on Matrix Analysis and Applications*, 31:2754–2768, 2010.
7. Christopher C. Paige, Beresford N. Parlett, and Henk A. van der Vorst. Approximate solutions and eigenvalue bounds from Krylov subspaces. *Numerical Linear Algebra with Applications*, 2:115–134, 1995.
8. Christopher C. Paige and Michael A. Saunders. Solution of sparse indefinite systems of linear equations. *SIAM Journal on Numerical Analysis*, 12:617–629, 1975.
9. Yousef Saad. *Iterative Methods for Sparse Linear Systems*. The PWS Publishing Company, Boston, 1996. Second edition, SIAM, Philadelphia, 2003.
10. Marcus Sarkis and Daniel B. Szyld. Optimal left and right additive Schwarz preconditioning for minimal residual methods with Euclidean and energy norms. *Computer Methods in Applied Mechanics and Engineering*, 196:1612–1621, 2007.
11. David J. Silvester and Valeria Simoncini. An optimal iterative solver for symmetric indefinite systems stemming from mixed approximation. *ACM Transactions on Mathematical Software*, 37:42:1–42:22, 2011.
12. Valeria Simoncini and Daniel B. Szyld. On the occurrence of superlinear convergence of exact and inexact Krylov subspace methods. *SIAM Review*, 47:247–272, 2005.
13. Valeria Simoncini and Daniel B. Szyld. Recent computational developments in Krylov subspace methods for linear systems. *Numerical Linear Algebra with Applications*, 14:1–59, 2007.
14. Daniel B. Szyld and Olof B. Widlund. Variational analysis of some Conjugate Gradient methods. *East-West Journal of Numerical Mathematics*, 1:51–74, 1993.
15. Abraham van der Sluis and Henk A. van der Vorst. The rate of convergence of Conjugate Gradients. *Numerische Mathematik*, 48:543–560, 1986.
16. Henk A. van der Vorst. *Iterative Krylov Methods for Large Linear Systems*. Cambridge University Press, Cambridge, New York, and Melbourne, 2003.
17. Henk A. van der Vorst and C. (Kees) Vuik. The superlinear convergence behaviour of GMRES. *Journal of Computational and Applied Mathematics*, 48:327–341, 1993.
18. Andrew J. Wathen, Bernd Fischer, and David J. Silvester. The convergence rate of the minimal residual method for the Stokes problem. *Numerische Mathematik*, 71:121–134, 1995.

Fluid-Structure Interaction: Acceleration of Strong Coupling by Preconditioning of the Fixed-Point Iteration

M.R. Dörfel and B. Simeon

Abstract This contribution focuses on partitioned solution approaches in fluid-structure interaction problems. Depending on certain physical parameters of fluid and structure, the fixed-point iteration that is mostly used to strongly couple the different solvers in each time step is susceptible to deceleration. We present a method that is able to overcome this effect by a specific preconditioning of the fixed-point iteration. Thus, the full convergence order of the underlying time-discretisation schemes is preserved. As computational example, a benchmark problem from hemodynamics is considered where this effect has a particularly strong influence. It turns out that, though a single step of the preconditioned iteration is more expensive, the overall gain in efficiency can be significant.

1 Introduction

Fluid-structure interaction (FSI) arises in various application fields such as the flutter analysis of wings of aircrafts [3] and hemodynamics [2]. The easiest and oldest solution approach for FSI problems is the so-called weak coupling or staggered scheme where in each time step, the structural and the fluid equations are solved separately and thereafter, boundary conditions are exchanged and the integration proceeds. Depending on the different properties of the involved materials, this partitioned scheme, however, is prone for instabilities [1, 7, 8].

M.R. Dörfel (✉)
Technische Universität München, Chair of Numerical Analysis, Boltzmannstr. 3, 85748, Garching, Germany
e-mail: doerfel@ma.tum.de

B. Simeon
Department of Mathematics, Technische Universität Kaiserslautern, Gottlieb-Daimler-Str. 47, 67663, Kaiserslautern, Germany
e-mail: simeon@mathematik.uni-kl.de

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_78,

A way out is the strong coupling of fluid and structure by enforcing the equations on the boundaries implicitly, which can be established in two different ways. The first possibility is to employ a monolithic approach where the equations for the fluid and the structure are solved simultaneously in a large system of nonlinear equations. To deal with the increase in system dimension, which is the main bottleneck of this approach, specific preconditioners in the linear algebra are necessary that take the block structure of the equations into account [1, 2]. An alternative, and today the most widely used approach, are strongly coupled partitioned schemes where the fields are still solved seperately and the implicit coupling at the boundary is performed by means of either Newton-type algorithms or a straightforward fixed-point iteration. Both options have their pros and cons. For example Newton-type methods, the required Jacobian is usually computed with finite differences and thus very expensive. On the other hand, the fixed-point iteration might need a large number of iterations to achieve convergence or even diverge if no acceleration method like Aitken's relaxation [9] is used [10].

This contribution wants to shed new light on the last issue by analysing when and why these convergence problems in the fixed-point iteration occur and how they can be avoided. For this purpose, we take up the ideas of [6] where a one-dimensional model problem has been studied. Major changes are proposed to extend it to the non-linear three-dimensional models that are typical in real-life applications. The method that we propose here is able to overcome the convergence problems by a specific preconditioning of the fixed-point iteration. Though a single step of this preconditioned iteration is more expensive, it turns out that the overall gain in efficiency can be significant. The outline of this contribution is as follows: In the following section, the problem setup is summarised, starting with the equations of non-linear elasticity and the Navier-Stokes equations in ALE form and concluding with the system discretised in space and time. In Sect. 3, Gauss-Seidel methods are presented which yield the fixed-point iterations that are analysed subsequently. Extending the idea of the mass shift presented in [6] to this general framework leads to the preconditioned iteration, which is introduced in Sect. 4. Finally, Sect. 5 contains numerical results for a benchmark problem from hemodynamics [7, 10] as well as a conclusion.

2 Problem Setup

In this section, the general framework for FSI problems is summarised. We omit some details and concentrate instead on the basic problem structure that is relevant for the subsequent analysis.

In the structural part of the domain, Ω_S, the Lagrangian view is employed, i.e. the basic unknown is the deformation $\mathbf{d} : \Omega_S \times \mathscr{I} \to \Omega_S^{\mathscr{I}}$ where $\mathscr{I}$ is the time-interval of interest and $\Omega_S^{\mathscr{I}}$ the space-time-tube comprising the deformed shapes $\Omega_S^t \times \{t\}$. This unknown is governed by the equations of non-linear elastodynamics

$$\rho_S \frac{\partial^2}{\partial t^2}\mathbf{d} - \operatorname{div}(\nabla \mathbf{d}\, \mathbf{S}) = \mathbf{0} \qquad \text{in } \Omega_S \times \mathscr{I} \tag{1}$$

where ρ_S is the density of the structure and $\mathbf{S}$ denotes the second Piola-Kirchhoff stress tensor, see, e.g., [4]. The velocity of the structure is denoted by $\mathbf{v} = \frac{\partial}{\partial t}\mathbf{d}$. For the fluid domain Ω_F, the Eulerian view is used where the velocity $\mathbf{u}$ and the pressure p are the basic variables. Due to the motion of the structure, the Arbitrary Lagrangian-Eulerian (ALE) view is employed here. This involves another variable, the grid deformation $\boldsymbol{\varphi} : \Omega_F \times \mathscr{I} \to \Omega_F^{\mathscr{I}}$, which is sometimes even refered to as an additional field, the ALE field. The governing equations for $\mathbf{u} : \Omega_F^{\mathscr{I}} \to \mathbb{R}^3$ and $p : \Omega_F^{\mathscr{I}} \to \mathbb{R}$ are the incompressible Navier-Stokes equations in ALE formulation

$$\rho_F \frac{\partial \mathbf{u}}{\partial t} + \rho_F((\mathbf{u} - \mathbf{u}^G) \cdot \nabla)\mathbf{u} - \operatorname{div} \boldsymbol{\sigma}_F = \mathbf{0} \quad \text{in } \Omega_F^{\mathscr{I}}, \quad \operatorname{div} \mathbf{u} = 0 \quad \text{in } \Omega_F^{\mathscr{I}} \tag{2}$$

where ρ_F is the density of the fluid, $\mathbf{u}^G$ is the grid velocity defined as $\mathbf{u}^G = \frac{\partial}{\partial t}\boldsymbol{\varphi}$, and $\boldsymbol{\varphi}$ is determined by pseudo-elasticity via

$$\operatorname{div}(k \nabla \boldsymbol{\varphi}) = 0 \quad \text{in } \Omega_F \times \mathscr{I} \tag{3}$$

where k is a stiffness constant. Additional to these field equations, coupling conditions are imposed on the common boundary $\Gamma_{\text{FSI}}^t = \Omega_S^t \cap \Omega_F^t$. They read

$$\mathbf{d} = \boldsymbol{\varphi}, \quad \mathbf{v} = \mathbf{u} = \mathbf{u}^G \quad \text{and} \quad \boldsymbol{\sigma}_S\, \mathbf{n}_S = \boldsymbol{\sigma}_F\, \mathbf{n}_F \qquad \text{on } \Gamma_{\text{FSI}}^t. \tag{4}$$

To discretise this coupled PDE system, the vertical method of lines is applied which starts with the spatial discretisation by means of the finite element method. Also, three out of the four coupling conditions in (4) are weakly enforced and coupled to the field equations by means of Lagrangian multipliers $\underline{\boldsymbol{\lambda}}_S$, $\underline{\boldsymbol{\lambda}}_F$ and $\underline{\boldsymbol{\lambda}}_A$. This step results in a system of differential-algebraic equations (DAE) for the time-dependent coefficient vectors. Note that the index of this DAE is not the same for different choices of (4). If we assume that the deformations of structure and ALE field are enforced by means of $\mathbf{d} = \boldsymbol{\varphi}$, the index is two for a coupling of the structure with the fluid velocities $\mathbf{v} = \mathbf{u}$ and three for a coupling of the ALE field with the fluid velocities $\mathbf{u}^G = \mathbf{u}$. The major effect of this difference is that the latter formulation is prone to result in instabilities if adaptive time stepping schemes are used, see [5] for more details. Because of this, the first coupling is chosen in the rest this article.

Next, a modular time discretisation with possibly different implicit, linear multistep methods is employed. To stay general, the following equations make use of a unifying notation that contains method-specific parameters given in Table 1. A comparison of the performance of the different time integration methods is omitted here but can be found in detail in [5]. Moreover, the terms depending on boundary conditions and the data of previous time steps (again specific for each method) are collected in $\boldsymbol{\Phi}_n^\star$ for each equation. We mark the discrete unknowns from above by

Table 1 Definitions of the method-specific parameters for the time discretisation

		BDF1	BDF2	Generalised α
Structure	$\tilde{\beta}$	1	$\frac{4}{9}$	$\beta_S \frac{1-\alpha_f}{1-\alpha_m}$
	$\tilde{\gamma}$	1	$\frac{2}{3}$	$\frac{\beta_S}{\gamma_S}$
Fluid	$\tilde{\theta}$	1	$\frac{2}{3}$	$\theta \frac{1-\delta_f}{1-\delta_m}$
ALE	$\tilde{\delta}$	1	$\frac{2}{3}$	$\frac{\beta_A}{\gamma_A}$

underlining and an additional index for the current time step, e.g., $\underline{\mathbf{d}}_{n+1} \doteq \underline{\mathbf{d}}(t_{n+1})$, and obtain finally the following non-linear system of purely algebraic equations

$$\mathbf{M}_S \underline{\mathbf{d}}_{n+1} + \tilde{\beta} \Delta t^2 \, \mathbf{k}_{S,n+1} - \tilde{\beta} \Delta t^2 \, \mathbf{C}_S \underline{\boldsymbol{\lambda}}_{S,n+1} = \boldsymbol{\Phi}_n^{S1} \tag{5a}$$

$$\underline{\mathbf{v}}_{n+1} - \frac{1}{\tilde{\gamma} \Delta t} \underline{\mathbf{d}}_{n+1} = \boldsymbol{\Phi}_n^{S2} \tag{5b}$$

$$\mathbf{C}_S^T \underline{\mathbf{d}}_{n+1} - \mathbf{C}_A^T \underline{\boldsymbol{\varphi}}_{n+1} = \mathbf{0} \tag{5c}$$

$$\mathbf{K}_A \underline{\boldsymbol{\varphi}}_{n+1} - \mathbf{C}_A \underline{\boldsymbol{\lambda}}_{A,n+1} = \boldsymbol{\Phi}_n^{A1} \tag{5d}$$

$$\underline{\mathbf{u}}_{G,n+1} - \frac{1}{\tilde{\delta} \Delta t} \underline{\boldsymbol{\varphi}}_{n+1} = \boldsymbol{\Phi}_n^{A2} \tag{5e}$$

$$\mathbf{C}_S^T \underline{\mathbf{v}}_{n+1} - \mathbf{C}_F^T \underline{\mathbf{u}}_{n+1} = \mathbf{0} \tag{5f}$$

$$\mathbf{M}_{F,n+1} \underline{\mathbf{u}}_{n+1} + \tilde{\theta} \Delta t \, \mathbf{k}_{F,n+1} - \tilde{\theta} \Delta t \, \mathbf{K}_{p,n+1} \underline{\mathbf{p}}_{n+1} - \tilde{\theta} \Delta t \, \mathbf{C}_F \underline{\boldsymbol{\lambda}}_{F,n+1} = \boldsymbol{\Phi}_n^F \tag{5g}$$

$$\mathbf{K}_{p,n+1}^T \underline{\mathbf{u}}_{n+1} + \mathbf{k}_{p,n+1} = \mathbf{0} \tag{5h}$$

$$\mathbf{C}_{n+1}^{\lambda S} \underline{\boldsymbol{\lambda}}_{S,n+1} + \mathbf{C}_{n+1}^{\lambda F} \underline{\boldsymbol{\lambda}}_{F,n+1} = \mathbf{0}. \tag{5i}$$

Here, standard notation is mostly used, i.e., $\mathbf{M}_\star$ are mass matrices, $\mathbf{K}_\star$ and $\mathbf{k}_\star$ are stiffness matrices and vectors, respectively, and $\mathbf{C}_\star$ stands for problem-specific coupling matrices. A subscript $n+1$ indicates a time-dependence of the corresponding terms. For the sake of simplicity, only the special case of 'matching grids' has been considered at the set-up of system (5).

3 Analysis of the Fixed-Point Iteration

As mentioned in the introduction, we consider the case of strong coupling where all equations of system (5) are enforced but the subproblems are still solved seperately. This implies that an iteration over the fields becomes necessary. This approach can also be interpreted as performing a Block-Gauss-Seidel iteration on the monolithic system (5) directly. Different methods arise by ordering the subsystems as well as the coupling equations in different orders.

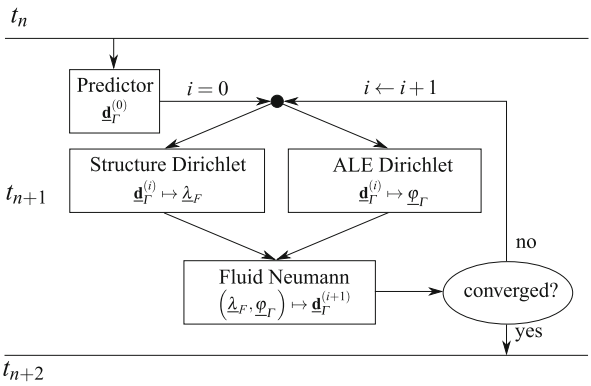


Fig. 1 The non-standard Dirichlet-Neumann algorithm omitting the indices $n+1$

It is standard to use the coupling equations in such a way that a Neumann problem is solved in the structure and a Dirichlet problem in the fluid and the ALE (in the ongoing refered to as the *standard iteration*). However, we focus here on an iteration where Dirichlet problems are defined for the structure and the ALE and a Neumann problem for the fluid. Both algorithms are members of the class of Dirichlet-Neumann iterations. The latter one is visualised in Fig. 1, which illustrates the necessity of a predictor and a convergence criterion. Like in the standard iteration, the predictor may be a coarse-mesh solution in a multigrid approach or simply the solution of an explicit time-step such as $\mathbf{d}_n + \mathbf{v}_n \Delta t$. Also the convergence criterion is the same as in the standard case and requires a small relative and/or absolute error of the increment $\mathbf{d}_\Gamma^{(i+1)} - \mathbf{d}_\Gamma^{(i)}$.

We next define the function $\mathscr{N}$ to describe a single step in the iteration

$$\underline{\mathbf{d}}_\Gamma^{(i+1)} = \mathscr{N}\left(\underline{\mathbf{d}}_\Gamma^{(i)}\right). \tag{6}$$

The convergence properties of this fixed-point iteration follow from the norm of the derivative of $\mathscr{N}$. Banach's fixed-point theorem states the uniqueness of a solution as well as convergence of (6) if $\mathscr{N}$ is a contraction, i.e. $\|\nabla\mathscr{N}\| < 1$. Linearising around the solution of the old time step t_n and solving the equations with Schur complements yields

$$\nabla\mathscr{N} = -\tau\, \mathbf{B}_F^{-1}\mathbf{B}_S + \mathbf{O}(\Delta t) \tag{7}$$

$$\text{with} \quad \mathbf{B}_S = \mathbf{G}_S^{\Gamma\Gamma} - \mathbf{G}_S^{\Gamma\mathrm{I}}\left(\mathbf{G}_S^{\mathrm{II}}\right)^{-1}\mathbf{G}_S^{\mathrm{I}\Gamma} \quad \text{and} \quad \mathbf{B}_F = \mathbf{G}_F^{\Gamma\Gamma} + \mathbf{A}_1(\mathbf{K}_{p,n}^{\Gamma})^T - \mathbf{A}_2\mathbf{G}_F^{\mathrm{I}\Gamma}$$

where $\mathbf{G}_S = \mathbf{M}_S + \tilde{\beta}\Delta t^2 \mathbf{y}\mathbf{K}_{S,n}$, $\mathbf{G}_{F,n} = \mathbf{M}_{F,n} + \tilde{\theta}\Delta t\ \mathbf{K}_{F,n}$ and with $\tau = \tilde{\gamma}\tilde{\theta}/\tilde{\beta}$ depending solely on the time integration constants. The matrices $\mathbf{A}_1$ as well as $\mathbf{A}_2$ comprise further Schur complements of the fluid field. Moreover, the indices Γ and I denote the set of rows and columns which correspond to the FSI-boundary and to the inner degree of freedoms, respectively. It follows for small Δt that $\mathbf{B}_S = \mathbf{O}(\rho_S)$

and $\mathbf{B}_F = \mathbf{O}(\rho_F)$, and thus $\|\nabla \mathcal{N}\| = O(\rho_S/\rho_F)$, i.e., there are parameters ρ_S and ρ_F for which (6) does not converge. This effect has been studied by several authors in the context of the standard iteration where it holds that $\nabla \hat{\mathcal{N}} = -1/\tau\ \mathbf{B}_S^{-1}\mathbf{B}_F + \mathbf{O}(\Delta t)$, resulting in the dependence of the convergence on the inverse ratio ρ_F/ρ_S, see [7, 8, 10].

A first and inexpensive possibility to improve the convergence is to employ a relaxation technique, most preferably a dynamic one like Aitken's relaxation [9]. This method is indeed able to prevent the fixed-point iteration from diverging, but still yields a very slow convergence if $\|\nabla \mathcal{N}\|$ is large [10]. A stronger acceleration method is therefore presented in the next section.

4 Preconditioning of the Iteration

The new acceleration method is best motivated by the method introduced in [6]. There, a similar iteration for a simplified FSI problem has been improved by incrementing the iteration index of certain terms in the Dirichlet problem and traversing this new unknown term through the iteration. Unfortunately, this is not feasible in the present context of a non-linear (and especially non-simplified) FSI problem. The same effect, however, can be achieved if the following alternative changes are applied. The main idea is to introduce an additional term $\mathbf{H}(\underline{\mathbf{d}}_\Gamma^{(i+1)} - \underline{\mathbf{d}}_\Gamma^{(i)})$ which goes to zero in the limit. To improve the convergence behaviour, this term is added to the equations which correspond to the boundary degree of freedoms of the structure. For the moment, $\mathbf{H}$ shall be an arbitrary matrix that will be specified later on as a preconditioner. Like in [6], the two unknowns $\underline{\boldsymbol{\lambda}}_S$ and $\underline{\mathbf{d}}_\Gamma^{(i+1)}$ are then connected to an auxiliary variable, but this time it holds that

$$\tilde{\underline{\boldsymbol{\lambda}}}_S = \underline{\boldsymbol{\lambda}}_S - \frac{1}{\tilde{\beta}\Delta t^2}\mathbf{A}_{\mathrm{FSI}}^{-1}\mathbf{H}\underline{\mathbf{d}}_\Gamma^{(i+1)} \tag{8}$$

with invertible matrix $\mathbf{A}_{\mathrm{FSI}} = \mathbf{C}_S^\Gamma = \mathbf{C}_A^\Gamma = \mathbf{C}_F^\Gamma$. In this way, the structural system becomes

$$\begin{pmatrix} \mathbf{M}_S^{\mathrm{II}} & \mathbf{M}_S^{\mathrm{I}\Gamma} \\ \mathbf{M}_S^{\Gamma\mathrm{I}} & \mathbf{y}\mathbf{M}_S^{\Gamma\Gamma} - \mathbf{H} \end{pmatrix} \begin{pmatrix} \underline{\mathbf{d}}_{\mathrm{I}} \\ \underline{\mathbf{d}}_\Gamma^{(i)} \end{pmatrix} + \tilde{\beta}\Delta t^2 \mathbf{k}_S\left(\underline{\mathbf{d}}_{\mathrm{I}}, \underline{\mathbf{d}}_\Gamma^{(i)}\right) - \tilde{\beta}\Delta t^2 \begin{pmatrix} \mathbf{0} \\ \mathbf{A}_{\mathrm{FSI}}\tilde{\underline{\boldsymbol{\lambda}}}_S \end{pmatrix} = \boldsymbol{\Phi}_n^{S1},$$

which has to be solved for $\underline{\mathbf{d}}_{\mathrm{I}}$ and $\tilde{\underline{\boldsymbol{\lambda}}}_S$. The latter is transferred to the fluid subsystem via (5i) like the standard Lagrangian multiplier giving $\tilde{\underline{\boldsymbol{\lambda}}}_F$. This unknown is then substituted in the following relations

$$\begin{aligned} \tilde{\theta}\Delta t\mathbf{A}_{\mathrm{FSI}}\tilde{\underline{\boldsymbol{\lambda}}}_F &= \tilde{\theta}\Delta t\mathbf{A}_{\mathrm{FSI}}\tilde{\underline{\boldsymbol{\lambda}}}_F - \frac{\tilde{\theta}\tilde{\gamma}}{\tilde{\beta}}\mathbf{A}_{\mathrm{FSI}}\left(\mathbf{C}^{\lambda F}\right)^{-1}\mathbf{C}^{\lambda S}\mathbf{A}_{\mathrm{FSI}}^{-1}\mathbf{H}\frac{1}{\tilde{\gamma}\Delta t}\underline{\mathbf{d}}_\Gamma^{(i+1)} \\ &= \tilde{\theta}\Delta t\mathbf{A}_{\mathrm{FSI}}\tilde{\underline{\boldsymbol{\lambda}}}_F - \mathbf{C}\mathbf{H}\,\underline{\mathbf{u}}_\Gamma^{(i+1)} + \mathbf{C}\mathbf{H}\,\boldsymbol{\Phi}_n^{S2,\Gamma}, \end{aligned}$$

where (5b) and (5f) and an according definition of $\mathbf{C}$ have been used in the last step. Inserting this result into the fluid subsystem yields, instead of (5g),

$$\begin{pmatrix} \mathbf{M}_F^{\mathrm{II}} & \mathbf{M}_F^{\mathrm{I}\Gamma} \\ \mathbf{M}_F^{\Gamma\mathrm{I}} & \mathbf{M}_F^{\Gamma\Gamma} + \mathbf{CH} \end{pmatrix} \begin{pmatrix} \underline{\mathbf{u}}_{\mathrm{I}} \\ \underline{\mathbf{u}}_{\Gamma}^{(i+1)} \end{pmatrix} + \tilde{\theta}\Delta t\ \mathbf{k}_F \left(\underline{\mathbf{u}}_{\mathrm{I}}, \underline{\mathbf{u}}_{\Gamma}^{(i+1)} \right) - \tilde{\theta}\Delta t\ \mathbf{K}_{p,n+1} \underline{\mathbf{p}}_{n+1} - \begin{pmatrix} \mathbf{0} \\ \tilde{\theta}\Delta t\ \mathbf{A}_{\mathrm{FSI}} \tilde{\underline{\boldsymbol{\lambda}}}_F \end{pmatrix} = \boldsymbol{\Phi}_n^F + \begin{pmatrix} \mathbf{0} \\ \mathbf{CH}\ \boldsymbol{\Phi}_n^{S2,\Gamma} \end{pmatrix},$$

which is solved together with (5h) for $\underline{\mathbf{u}}_{\mathrm{I}}$, $\underline{\mathbf{u}}_{\Gamma}^{(i+1)}$ and $\underline{\mathbf{p}}_{n+1}$.

Note that the proposed changes just require the mass matrices $\mathbf{M}_S$ and $\mathbf{M}_F$ and the right-hand side of the fluid subsystem $\boldsymbol{\Phi}_n^F$ to be altered while the standard solver can be applied. It solves then for $\tilde{\underline{\boldsymbol{\lambda}}}_{S/F}$ instead of $\underline{\boldsymbol{\lambda}}_{S/F}$. If the latter are needed, they can be evaluated after convergence using (8).

This accelerated method indeed converges to the same solution as the original iteration [5]. Abbreviating it as $\underline{\mathbf{d}}_{\Gamma}^{(i+1)} = \mathscr{N}_{\mathbf{H}} \left(\underline{\mathbf{d}}_{\Gamma}^{(i)} \right)$, it satisfies

$$\nabla \mathscr{N}_{\mathbf{H}} = -\tau\ (\mathbf{B}_F + \mathbf{CH})^{-1} (\mathbf{B}_S - \mathbf{H}) + \mathbf{O}(\Delta t),$$

which leads to an optimal convergence if $\mathbf{H} = \mathbf{B}_S$ is chosen.

5 Numerical Example and Conclusion

The following example goes back to [7, 10] and is widely used as a benchmark problem for FSI in hemodynamics. It consists of a three dimensional flexible tube of length $l = 5\,\mathrm{cm}$, inner radius $r_1 = 0.5\,\mathrm{cm}$, and outer radius $r_2 = 0.6\,\mathrm{cm}$. The structure is an isotropic St. Venant-Kirchhoff material with Young's modulus $E = 3 \cdot 10^5\,\mathrm{Pa}$, Poisson ratio $\nu = 0.3$ and density $\rho_S = 1,200\,\mathrm{kg/m^3}$. The fluid is characterised by its density $\rho_F = 1,000\,\mathrm{kg/m^3}$ and viscosity $\mu_F = 0.003\,\mathrm{Pa\,s}$. At the inflow, a pressure wave is initiated at $t = 0$ which is traversing through the tube as can be seen in Fig. 2. For these material constants, both the standard and the non-standard Dirichlet-Neumann algorithm diverge without additional measures. Using Aitken's relaxation, both iterations converge but require a substantial number of iterations, see Table 2. This table also compares the total used time as well as the time per iteration which are both given in relation to the standard algorithm. Whilst all iterations converge to the same solution that has also been validated with the black box solver COMSOL, we observe that the preconditioned algorithm is indeed outperforming the other options. The last column shows, however, that the dense preconditioner $\mathbf{H} = \mathbf{B}_S$ slows down the solution of the non-linear systems within one iteration. This indicates the potential of a further improvement of the method by using a sparse preconditioner instead of $\mathbf{B}_S$. The same modification is also necessary

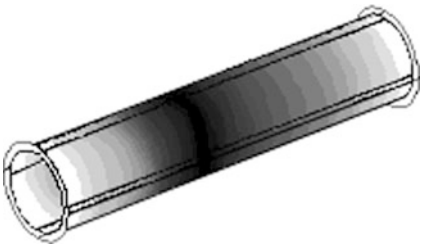


Fig. 2 The displacement of the FSI boundary at $t = 0.0055$ s – in total approximately 150,000 degrees of freedom are used

Table 2 Comparison of the algorithms using $\Delta t = 10^{-4}$ s, $T = 0.001$ s, and a tolerance for the relative displacements of 10^{-5} as convergence criteria

	Average number of iterations	Total relative time	Relative time per iteration
Standard iteration	51.7	1.00	1.00
Non-standard iteration of Fig. 1	204.7	4.29	1.08
Preconditioned iteration	2.0	0.15	3.42

in case of an increased system size where it is not recommendable to calculate the expression with the inverse $\mathbf{G}_S^{\Gamma \mathrm{I}} \left(\mathbf{G}_S^{\mathrm{II}}\right)^{-1} \mathbf{G}_S^{\mathrm{I}\Gamma}$ within $\mathbf{B}_S$ explicitly.

In conclusion, further research in the direction of preconditioners for the fixed-point iteration appears to be very promising. Another option is an algorithm where the fluid system is only solved once whereas the computationally cheaper structural and ALE subsystems are solved twice to correct the additional term $\mathbf{H}(\underline{\mathbf{d}}_\Gamma^{(1)} - \underline{\mathbf{d}}_\Gamma^{(0)})$ or the use of the inaccurate $\underline{\mathbf{d}}_\Gamma^{(0)}$, respectively. This simplified algorithm would define a quasi-weak algorithm featuring the distinguishing mark of a weak coupling that no convergence test for the iteration can be applied. First results indicate that in this way another 30 % of the computing time can be saved without decreasing the accuracy.

Acknowledgements The first author was supported by Deutsche Forschungsgemeinschaft (DFG) through the TUM International Graduate School of Science and Engineering (IGSSE) within Project 2-11. This support is gratefully acknowledged.

References

1. Badia, S., Quaini, A., Quarteroni, A.: Modular vs. non-modular preconditioners for fluid-structure systems with large added-mass effect. Comput. Meth. Appl. Mech. Eng. **197**, pp. 4216–4232 (2008)
2. Bazilevs, Y., Calo, V. M., Zhang, Y., Hughes, T.J.R.: Isogeometric Fluid-structure interaction analysis with applications to arterial blood flow. Comput. Mech. **38**, pp. 310–322 (2006)
3. Bungartz, H.J., Schäfer, M.: Fluid-Structure Interaction, Modelling, Simulation, Optimisation. In: Lecture Notes in Computational Science and Engineering, Vol. 53, Springer (2006)
4. Ciarlet, P.G.: Mathematical Elasticity - Volume 1: Three-Dimensional Elasticity, North-Holland (1988)

5. Dörfel, M.R.: Fluid-Structure Interaction: A Differential-Algebraic Approach and Acceleration Techniques for Strong Coupling, In: Fortschritt-Berichte VDI Reihe 20, No. 436, VDI Verlag (2011)
6. Dörfel, M.R., Simeon, B.: Analysis and acceleration of a fluid-structure interaction coupling scheme. In: Kreiss, G. et al. (Eds.) Numerical Mathematics and Advanced Applications 2009, pp. 307–316. Springer (2010)
7. Fernndez, M.A., Gerbeau, J.-F., Grandmont, C.: A projection semi-implicit scheme for the coupling of an elastic structure with an incompressible fluid. Technical Report INRIA Rocquencourt **5700**, pp.1–31 (2005)
8. Förster, C., Wall, W.A., Ramm, E.: Artificial added mass instabilities in sequential staggered coupling of nonlinear structures and incompressible viscous flows. Comput. Meth. Appl. Mech. Eng. **196**, pp. 1278–1293 (2007)
9. Irons, B., Tuck, R.C.: A version of the Aitken accelerator for computer implementation. Int. J. Numer. Meth. Eng. **1**, pp. 275–277 (1969)
10. Küttler, U., Wall, W.A.: Fixed-point fluid-structure interaction solvers with dynamic relaxation. Comput. Mech. **43**, pp. 61–72 (2008)

Some Experiences with Multilevel Krylov Methods

Y.A. Erlangga

Abstract This note discusses convergence behaviors of multilevel Krylov methods for some simple problems, mainly focusing on the possible choice of transfer operators. This study is part of the search for an optimal multilevel Krylov method.

1 Introduction

In [3], Erlangga and Nabben proposed a multilevel method for solving the linear system

$$Ax = b, \quad A \in \mathbb{R}^{n \times n}, \tag{1}$$

which mimics a multigrid process, but, instead of a smoother, with a Krylov method used at each level. Since a (optimal) Krylov method reduces, not smoothes, errors in some norm, the underlying concept of this Krylov-based multilevel method is different from that of multigrid. Algorithm 4 shows the main body of the two-level version of a multilevel Krylov (MK) method, based on (flexible) GMRES [7].

Notice that a multigrid component is in play: the course-grid solve, which involves a coarse-grid matrix A_c. In multigrid, this solve is associated with the reduction of slow varying components of errors. The fine and coarse subspace are connected by two transfer operators I_H^h and I_h^H, which prolongate and restrict, respectively, some quantities in the iteration process. In multigrid, these quantities are related to errors made by an approximate solution, obtained after a few smoothing steps or coarse-grid solves. In MK, this quantity is associated with vectors, which build the approximation subspace.

Y.A. Erlangga (✉)
Mathematics Department, Alfaisal University, At-Takhassusi Road (South), P.O. Box 10927, Riyadh, 11533, Saudi Arabia
e-mail: yerlangga@alfaisal.edu; yogiae@gmail.com

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_79,

Algorithm 4 Two-level Krylov

Set an initial guess of solution x_0;
Compute the residual $r_0 = b - Ax_0$; Set $v_1 = r_0/\|r_0\|$;
For $j = 1, \ldots,$ until convergence
  $p = Av_j$;
  $q = I_h^H(p - \lambda_n v_j)$;
  Solve for s: $A_c s = q$;
  $z_j = v_j - I_H^h s$;
  $w = Az_j$;
  Gram-Schmidt orthogonalization;
End
Extract the approximate solution from $Z = [z_1 \ldots z_j]$.

From an abstract point of view, if the second-level problem is solved exactly, Algorithm 4 is the consequence of applying GMRES on the preconditioned linear system

$$AQ_N\tilde{x} = b, \quad x = Q_N\tilde{x}, \tag{2}$$

with $Q_N = I - I_H^h A_c^{-1} I_h^H A + \lambda_n I_H^h A_c^{-1} I_h^H$ and $\lambda_n = \max\{|\lambda_i|\}_{1\le i\le n}$, with λ_i the eigenvalue of A. Q_N is called the shift operator, due to the following spectral property [3].

Theorem 1. *Let A be symmetric positive definite and $\sigma(A) = \{\lambda_1, \ldots, \lambda_n\}$, $0 < \lambda_i \le \lambda_j$, for $i < j$. If $I_H^h = [z_1 \ldots z_m]$, with $Az_i = \lambda_i z_i$, $m < n$, $I_h^H = (I_H^h)^T$, and $A_c = (I_H^h)^T A I_H^h$ (Galerkin coarse grid), then $\sigma(AQ_N) = \{\lambda_{m+1}, \ldots, \lambda_n\}$, with λ_n having multiplicity of at least $m + 1$.*

Thus, under the assumptions set in Theorem 1, Q_N somehow shifts m smallest eigenvalues of A to λ_n, without changing the rest, making the spectrum more clustered. Obviously, the system (2) is more favorable for a Krylov method than the original system (1).

The first two terms in Q_N form the (right) deflation operator [6], denoted by Q_D. The effect of Q_N and Q_D on A are spectrally equivalent in the following sense [3].

Theorem 2. *Let A be symmetric positive definite. Let $Q_D = I - I_H^h A_c^{-1} I_h^H A$.*

$$\sigma(AQ_D) := \{0, \mu_{m+1}, \ldots, \mu_n\} \iff \{\mu_{m+1}, \ldots, \mu_n, \lambda_n\} =: \sigma(AQ_N).$$

Furthermore, if $P_D = Q_D^T$ and $P_N = I - AI_H^h A_c^{-1} I_h^H + \lambda_n I_H^h A_c^{-1} I_h^H$, then $\sigma(P_D A) = \sigma(AQ_D)$ and $\sigma(P_N A) = \sigma(AQ_N)$, and the same equivalence holds.

Theorem 2 can be easily extended to a more general class of matrices A.

In this short note, we present observed convergence behaviors of this method, based on some relatively simple problems. We shall base the presentation on the two level Krylov method (Algorithm 4), which represents the best performance possibly attained in terms of numbers of iterations, for a given multilevel Krylov setup.

Throughout this note, $\sigma(\cdot)$, $\lambda(\cdot)$, $\kappa(\cdot)$, $\mathscr{R}(\cdot)$, and $\mathscr{N}(\cdot)$ denote respectively the spectrum, eigenvalue, condition number, range, and null space of the argument.

2 Spectral Properties and Observed Convergence

In this section, we shall consider a hypothetical problem: a diagonal matrix $A = diag(1, 2, \ldots, n)$. In this case, $\sigma(A) = \{1, 2, \ldots, n\}$, the eigenvectors $v_i = \alpha_i \mathbf{e}_i$, $\alpha_i \neq 0$, and for any $b = [b_i]$, the linear system (1) has the solution $x = [x_i]_{1 \leq i \leq n} = [b_i / a_{ii}]$.

2.1 Eigenvectors

Let $I_H^h = [\mathbf{e}_i]_{1 \leq i \leq m}$, $I_h^H = (I_H^h)^T$, and set $A_c = I_h^H A I_H^h$ (Galerkin-type coarse-grid matrix). In this case, according to Theorem 2, $\sigma(P_N A) = \{m + 1, \ldots, n\}$, with $\lambda_n = n$ having a multiplicity $m + 1$. Furthermore, $\sigma(P_D A) = \{0, m + 1, \ldots, n\}$. As $P_D A$ is symmetric positive semi-definite, $\kappa_{\text{eff}}(P_D A) = n/(m + 1) < n = \kappa(A)$.

Let $x_0 = 0$, and consider the left-preconditioned version of the two-level Krylov method: $P_N A x = P_N b$. The Krylov subspace after the k-th iteration is

$$\mathscr{K}(P_N A, r_0) = \text{span}\{r_0, P_N A r_0, \ldots, (P_N A)^{k-1} r_0\}.$$

Straight-forward computations lead to the following results:

$$r_0 = \lambda_n D_m^{-1} b + D_{n-m}^0 b, \quad (P_N A)^i r_0 = \lambda_n^i D_m^{-1} b + D_{n-m}^{i-1} b, \ i = 1, \ldots, k-1,$$

with $D_m^{-1} = diag(A_m^{-1}\ 0)$ and $D_{n-m} = diag(0\ A_{n-m})$. Thus,

$$\begin{aligned} x_{k,N} &\in \text{span}\{\lambda_n D_m^{-1} b + D_{n-m}^0 b, \lambda_n D_m^{-1} b + D_{n-m} b, \ldots, \lambda_n^{k-1} D_m^{-1} b + D_{n-m}^{k-1} b\} \\ &\subseteq \text{span}\{D_m^{-1} b, D_{n-m}^0 b, D_{n-m} b, \ldots, D_{n-m}^{k-1} b\}. \end{aligned}$$

Consider the (deflated) CG iteration applied to $P_D A x = P_D b$. In this case, CG generates a sequence of approximate solutions, $\{\tilde{x}_{k,D}\}$, to this singular linear system such that

$$\tilde{x}_{k,D} \in \text{span}\{D_{n-m}^0 b, D_{n-m} b, \ldots, D_{n-m}^{k-1} b\}.$$

The solution of $Ax = b$ is then constructed as follows: $x = (I - P_D^T)x + P_D^T x = I_H^h A_c^{-1} I_h^H b + P_D^T x$, implying a sequence $x_{k,D} = I s_H^h A_c^{-1} I_h^H b + P_D^T \tilde{x}_{k,D}$, and

$$\begin{aligned} x_{k,D} &\in I_H^h A_c^{-1} I_h^H b + \text{span}\{P_D^T D_{n-m}^0 b, P_D^T D_{n-m} b, \ldots, P_D^T D_{n-m}^{k-1} b\} \\ &= D_{n-m}^{-1} b + \text{span}\{D_{n-m}^0 b, D_{n-m} b, \ldots, D_{n-m}^{k-1} b\}. \end{aligned}$$

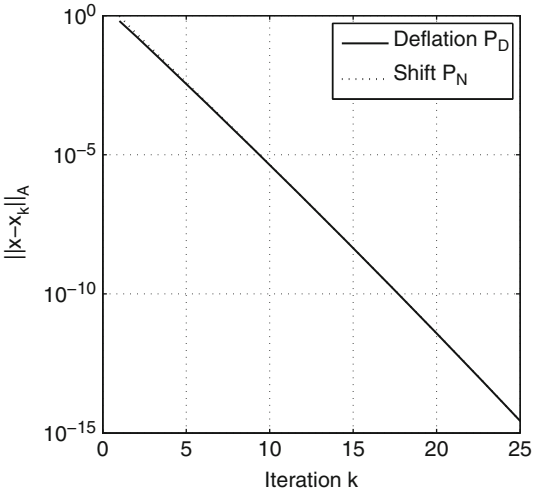


Fig. 1 Convergence history of CG with deflation P_D (*solid*) and shift P_N (*dotted*)

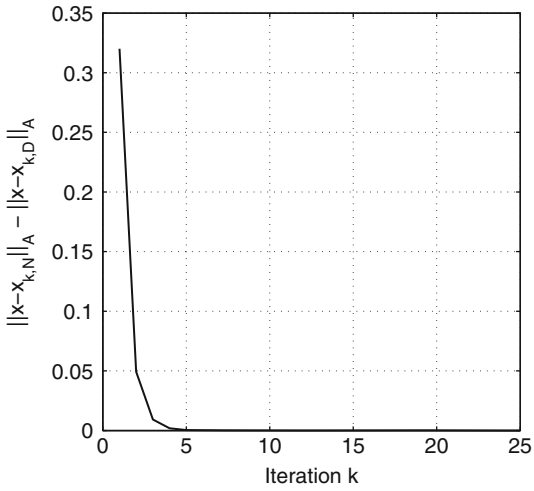


Fig. 2 Difference in the A-norm: $\|x - x_{k,N}\|_A - \|x - x_s k, D\|_A$

So, $x_{k,N}$ and $x_{k,D}$ are members of the same subspace. From [5], for every $x_k \in \text{span}\{D_{n-m}^{-1}b, D_{n-m}^{0}b, D_{n-m}b, \ldots, D_{n-m}^{k-1}b\} \equiv \mathscr{L}$,

$$\|x - x_{k,D}\|_A^2 \leq \|x - x_k\|_A^2.$$

Since $x_{k,N} \in \mathscr{L}$,

$$\|x - x_{k,D}\|_A^2 \leq \|x - x_{k,N}\|_A^2.$$

This result suggests the superiority of deflation technique to the shift operator, shown in Figs. 1 and 2 for the hypothetical problem.

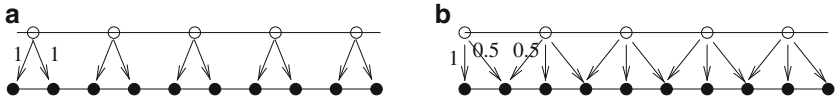


Fig. 3 One dimensional (**a**) aggregation and (**b**) linear interpolation, with ∘ and • indicating the fine and course nodes, respectively

A similar result holds also for (right preconditioned) GMRES, reading

$$\|b - Ax_{k,D}\|_2 \leq \|b - Ax_{k,N}\|_2.$$

The proof is skipped, and will be presented in another paper.

2.2 Interpolation Matrices

Consider any sparse, full rank interpolation matrix $I_H^h \in \mathbb{R}^{n\times m_1}$ and let $\mathscr{I}_H^h \in \mathbb{R}^{n\times m_2}$ be another full rank matrix. The following theorem is proved in [5]:

Theorem 3. *Let* $A_{c,1} = (I_H^h)^T A I_H^h$ *and* $A_{c,2} = (\mathscr{I}_H^h)^T A \mathscr{I}_H^h$. *Let* $P_{D,1} = I - AI_H^h A_{c,1}^{-1}(I_H^h)^T$, *and similarly for* $P_{D,2}$. *If* $\mathscr{R}(I_H^h) \subseteq \mathscr{R}(\mathscr{I}_H^h)$, *then*

1. $\lambda_n(P_{D,1}A) \geq \lambda_n(P_{D,2}A)$, *and*
2. $\lambda_{m_1+1}(P_{D,1}A) \leq \lambda_{m_2+1}(P_{D,2}A)$.

The *classical* examples of interpolation matrices are those associated with aggregation and linear interpolation in multigrid. They are illustrated for 1D in Fig. 3. Let I_H^h and $\mathscr{I}_H^h$ be matrices, associated with the "standard" aggregation and augmented aggregation, respectively. For the latter, we augment I_H^h by a column vector $(1\,0\,\ldots\,0\,1)^T$. So, $m_2 = m_1 + 1$, I_H^h and $\mathscr{I}_H^h$ are full rank, and $I_H^h \subseteq \mathscr{I}_H^h$. Thus, Theorem 3 holds. But, $P_N A$ (or AQ_N) is no longer symmetric, even if A is a diagonal matrix; CG certainly breaks down in this case, and GMRES has to be employed. Convergence of two-level Krylov is shown in Fig. 4. The figure suggests the faster convergence of the method with augmented aggregation. For deflation, this performance is predicted by Theorem 3 and the well-known convergence bound of CG (due to $\kappa_{\text{eff}}(P_{D,\mathscr{I}}A) \leq \kappa_{\text{eff}}(P_{D,I}A)$). For P_N, the GMRES convergence bound of [7] is, however, not useful for extracting detailed information about the behavior of the method (see the residuals at the initial stage of iterations). However, better clustering affects the overall convergence.

In Fig. 4, we show also the convergence based on the linear interpolation. The associated interpolation matrix (denoted by $\mathscr{I}_H^h$) is set such that it is of the same rank as the aggregation matrix (denoted by I_H^h). In this case, however, the inclusion condition of Theorem 3 does not hold. Use of linear interpolation clearly leads to a better convergence. This behavior is unfortunately not generally the rule, as we shall see in some examples in the subsequent sections.

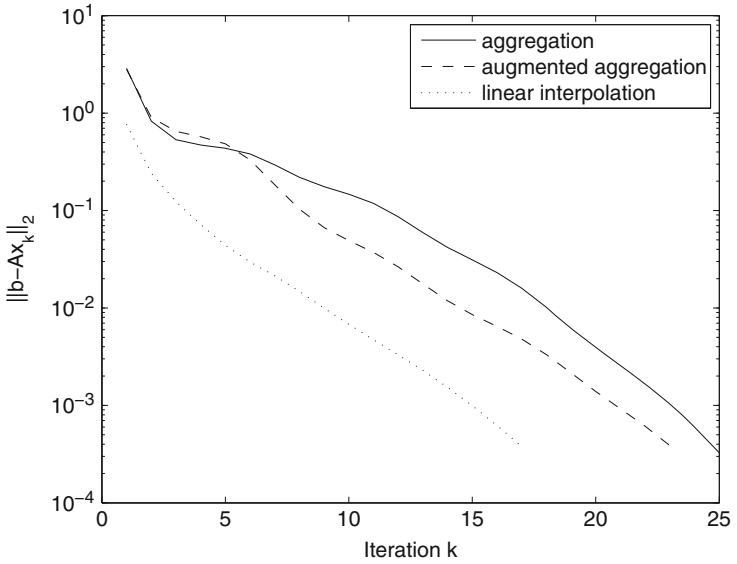


Fig. 4 GMRES convergence history, based on "standard" aggregation, augmented aggregation, and linear interpolation

3 Does Eigenvector-Based Transfer Operator Lead to a Better Convergence?

An almost general wisdom in deflation is to use eigenvectors or some (accurate) approximations to accelerate the convergence. The only reason one in practice avoids using them is because of the computational cost of computing even some of them. We shall address this issue in this section.

We consider a 1D Poisson problem in $[0, 1]$:

$$-u''(x) = f(x), \quad u(0) = u(1) = 0, \tag{3}$$

discretized by the second-order central difference. The eigenvalues and eigenvectors of the associated finite difference matrix A are

$$\lambda_k = 4\sin^2\left(\frac{\theta_k}{2}\right), \quad v_k = (\sin\theta_k, \sin(2\theta_k), \ldots, \sin(n\theta_k))^T, \quad k = 1, \ldots, n,$$

with $\theta_k = k\pi/(n+1)$, with n the number of interior grid points (and hence, the size of A). In Fig. 5, we compare the performance of two-level MK method based on eigenvectors, (augmented) aggregation, and linear interpolation for $n = 200$ and $m = n/2 = 100$. Interestingly, methods based on aggregation perform the best; they converge to the machine accuracy in two iterations. The spectrum of AQ_N in this case consists of just two eigenvalues, i.e., $\sigma(AQ_N) = \{2, 3.999\}$, while with eigenvectors, $\sigma(AQ_N) = \{100, 101, \cdots, 200\}$.

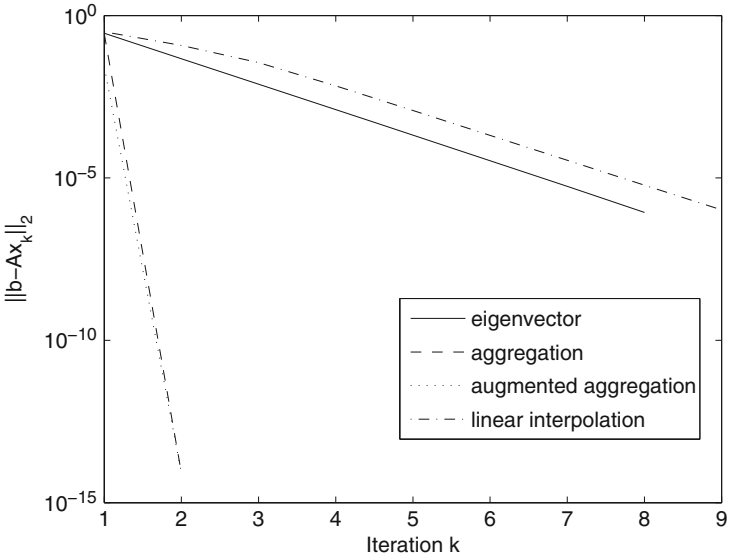


Fig. 5 GMRES convergence history for a 1D Poisson problem, based on eigenvectors, aggregation, and linear interpolation

Figure 6 shows a similar comparison, based on the 2D version of (3), discretized on a uniform finite difference mesh. In this case, eigenvector-based two-level method converges faster than the other scenarios, but aggregation-based techniques remain promising, and the (bi)linear interpolation performs the worst. This kind of performance is not one that we typically expect from multigrid (bilinear interpolation works well, and aggregation does not converge).

4 Singular but Consistent Systems

The last example is based on the 2D diffusion equation with Neumann conditions at the boundaries, set such that the resultant linear system is consistent. The matrix A is now singular, and is of rank $n-1$. If A_c is nonsingular, it can be shown that $\mathcal{N}(P_N A) = \mathcal{N}(AQ_N) = \mathcal{N}((P_N A)^T)$ [2]. In this case, according to Theorem 2.4 of [1], GMRES is guaranteed to converge to the least-squares solution of (2). An invertible coarse-grid matrix A_c can be obtained by modifying a_{nn} such that the sum of the last row is nonzero.

Figure 7 shows the convergence history for different choices of I_H^h, for constant density. In this case, the (augmented) aggregation strategy outperforms both eigenvector– and bilinear interpolation–based approach. Convergence of the problem with one bubble is shown in Fig. 8. Use of eigenvectors leads to the fastest convergence, even though it is not as fast as the convergence seen from the previous examples.

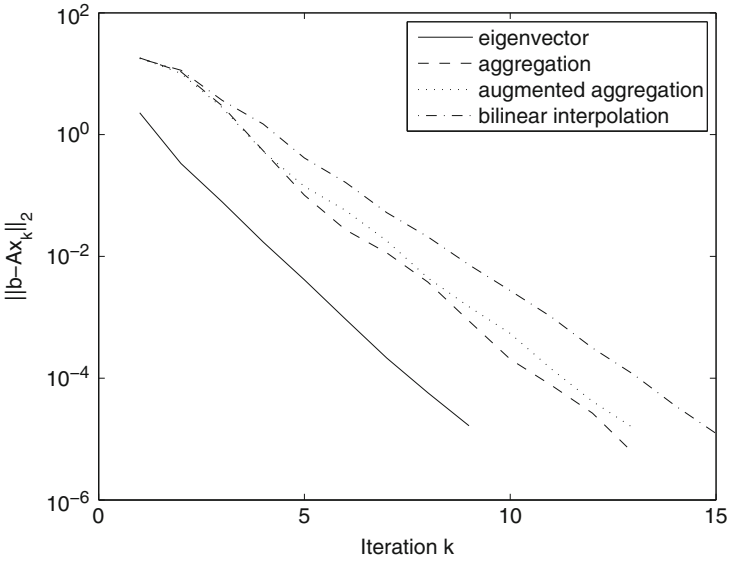


Fig. 6 GMRES convergence history for a 2D Poisson problem, based on eigenvectors, aggregation, and bilinear interpolation

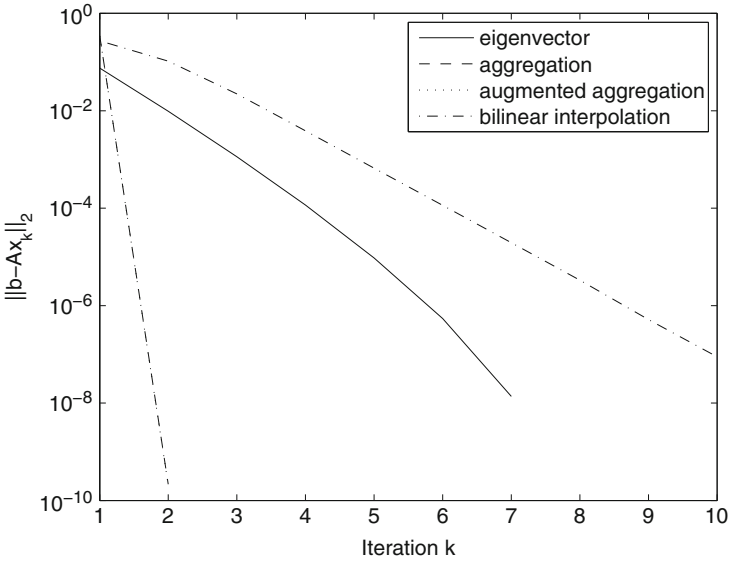


Fig. 7 GMRES convergence history for a 2D diffusion problem with uniform density

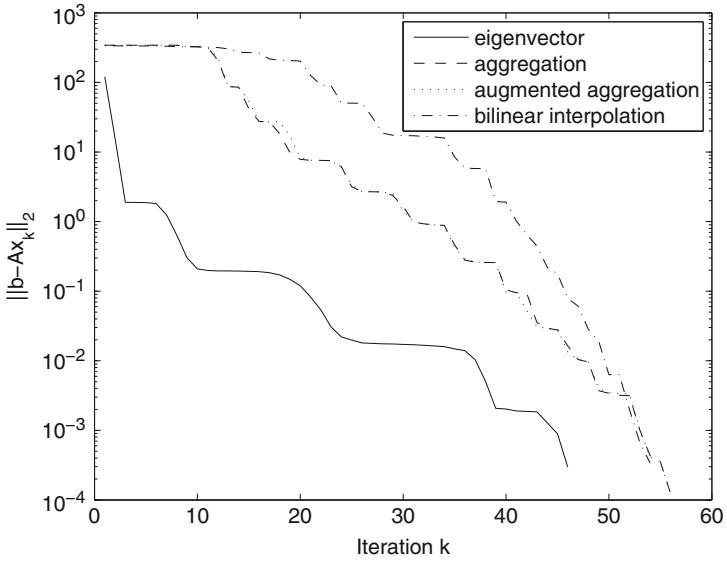


Fig. 8 GMRES convergence history for a 2D diffusion problem with one bubble and density ratio 10^3

5 Final Remarks

The potential of multilevel Krylov methods has been demonstrated in a number of papers; see also, e.g., [4]. This short note sheds some additional glimpse about this potential. Some examples suggest that eigenvector-based approach does not necessarily turn out to be the best way of setting up a multilevel Krylov method. Therefore, one can also argue that any approximation to eigenvectors may also lead to a non-optimal method. On the other hand, some examples show that it may be possible to get an extremely fast converging method using a rather simple transfer operator. What is crucial in this respect seems to lie in the choice of the transfer operator I_H^h. Unfortunately, at this point, no guideline is available in this direction.

References

1. Peter N Brown and Homer F Walker. GMRES on (nearly) singular systems. *SIAM J. Matrix Anal. Appl.*, 18:37–51, 1997.
2. Yogi A Erlangga. Multilevel krylov for singular systems. *manuscript*.
3. Yogi A Erlangga and Reinhard Nabben. Multilevel projection-based nested Krylov iteration for boundary value problems. *SIAM J. Sci. Comput.*, 30:1572–1595, 2008.
4. Yogi A Erlangga and Reinhard Nabben. Algebraic multilevel Krylov methods. *SIAM J. Sci. Comput.*, 31:3417–3437, 2009.

5. R Nabben and C Vuik. A comparison of deflation and coarse grid correction applied to porous media flow. *SIAM J. Numer. Anal.*, 42:1631–1647, 2004.
6. R A Nicolaides. Deflation of conjugate gradients with applications to boundary value problems. *SIAM J. Numer. Anal.*, 24:355–365, 1987.
7. Y Saad and M H Schultz. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. Stat. Comput.*, 7:856–869, 1986.

Preconditioning of Elasticity Problems with Discontinuous Material Parameters

I. Georgiev and J. Kraus

Abstract We consider preconditioning methods for the systems of linear algebraic equations arising from Symmetric Interior Penalty discontinuous Galerkin (SIPG) discretization of linear elasticity problems in primal (displacement) formulation. The presented approach is a generalization of the subspace correction method studied in Ayuso et al. (A Subspace correction method for discontinuous Galerkin discretizations of linear elasticity equations, arXiv:1110.5743v2, 2011) for linear elasticity problems with discontinuous material properties. The application of the preconditioner reduces to the solution of a problem arising from discretization of the equations of linear elasticity by nonconforming Crouzeix-Raviart finite elements plus the solution of a well-conditioned problem on the complementary space.

1 Introduction

Let the domain $\Omega \subset \mathbb{R}^2$ be a convex polygon and let $\boldsymbol{u}$ be a vector field in $\mathbb{R}^2$, defined on Ω such that $\boldsymbol{u} \in [H^2(\Omega)]^2$. Let $\Omega = \Omega_1 \cup \Omega_2$ be the union of two subdomains with different material properties. We denote by $\Gamma_{int} = \partial\Omega_1 \cap \partial\Omega_2$ the interface between the subdomains. The corresponding linear elasticity problem reads: find the displacement field $\boldsymbol{u}$ such that

I. Georgiev (✉)
Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Acad. G. Bonchev Str., Bl. 8, 1113 Sofia, Bulgaria

Johann Radon Institute for Computational and Applied Mathematics, Austrian Academy of Sciences, Altenbergerstraße 69, A-4040 Linz, Austria
e-mail: ivan.georgiev@oeaw.ac.at

J. Kraus
Johann Radon Institute for Computational and Applied Mathematics, Austrian Academy of Sciences, Altenbergerstraße 69, A-4040 Linz, Austria
e-mail: johannes.kraus@oeaw.ac.at

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_80,

$$
\begin{aligned}
-\mathrm{div}\boldsymbol{\sigma}(\boldsymbol{u}) &= \boldsymbol{f}, && \text{on } \Omega,\\
[[\sigma(\boldsymbol{u})\boldsymbol{n}]] &= \boldsymbol{0}, && \text{on } \Gamma_{int},\\
[[\boldsymbol{u}]] &= \boldsymbol{0}, && \text{on } \Gamma_{int},\\
\boldsymbol{u} &= \boldsymbol{0}, && \text{on } \Gamma_D,\\
\boldsymbol{\sigma}(\boldsymbol{u})\boldsymbol{n} &= \boldsymbol{g}, && \text{on } \Gamma_N.
\end{aligned}
\tag{1}
$$

Here $\boldsymbol{u}$ vanishes on Γ_D, which denotes the part of the boundary $\partial\Omega$ where Dirichlet boundary conditions are imposed, and the normal stresses are prescribed on the rest of the boundary Γ_N where $\boldsymbol{n}$ is the outward unit normal vector on $\partial\Omega$. The stress tensor $\boldsymbol{\sigma}$ is related to the displacements $\boldsymbol{u}$ by Hooke's law

$$\boldsymbol{\sigma}(\boldsymbol{u}) = \lambda_i trace(\boldsymbol{\varepsilon}(\boldsymbol{u}))\boldsymbol{I} + 2\mu_i\boldsymbol{\varepsilon}(\boldsymbol{u}) \quad \text{on } \Omega_i \quad i = 1, 2,$$

where λ_i and μ_i are the Lamé constants corresponding to the subdomains Ω_i and

$$\boldsymbol{\varepsilon}(\boldsymbol{u}) = \frac{1}{2}(\nabla\boldsymbol{u} + (\nabla\boldsymbol{u})^T)$$

is the symmetric part of the gradient of $\boldsymbol{u}$.

Let $\mathscr{T}_h$ be a shape-regular triangulation of Ω aligned with subdomains Ω_1 and Ω_2. We denote by h_T the diameter of the triangle T and we set $h = \max_{T\in\mathscr{T}_h} h_T$. A face shared by two neighboring elements (or being part of the boundary) is denoted by E. The set of all faces we will denote by $\mathscr{E}_h$, and the collection of all interior and boundary faces by $\mathscr{E}_h^o$ and $\mathscr{E}_h^\partial$. Further, the set of Dirichlet faces is denoted by $\mathscr{E}_h^D$, and the set of Neumann faces by $\mathscr{E}_h^N$. We thus have,

$$\mathscr{E}_h = \mathscr{E}_h^o \cup \mathscr{E}_h^\partial, \quad \mathscr{E}_h^D = \mathscr{E}_h^\partial \cap \Gamma_D, \quad \mathscr{E}_h^N = \mathscr{E}_h^\partial \cap \Gamma_N, \quad \mathscr{E}_h^\partial = \mathscr{E}_h^D \cup \mathscr{E}_h^N.$$

For two vector fields $\boldsymbol{v}$ and $\boldsymbol{w}$, which are sufficiently smooth so that the integrals below exist, we denote

$$(\boldsymbol{v},\boldsymbol{w})_{\mathscr{T}_h} = \sum_{T\in\mathscr{T}_h}\int_T \langle\boldsymbol{v},\boldsymbol{w}\rangle, \quad (\boldsymbol{v},\boldsymbol{w})_{\mathscr{E}} = \sum_{E\in\mathscr{E}}\int_E \langle\boldsymbol{v},\boldsymbol{w}\rangle,$$

where $\langle\cdot,\cdot\rangle$ and $\langle\cdot:\cdot\rangle$ are the Euclidean and the Frobenius inner products,

$$\langle\boldsymbol{v},\boldsymbol{w}\rangle = \sum_{k=1}^{2} v_k w_k, \qquad \langle\boldsymbol{v}:\boldsymbol{w}\rangle = \sum_{j=1}^{2}\sum_{k=1}^{2} v_{jk} w_{jk}.$$

To define the average and jump trace operators for an interior face $E \in \mathscr{E}_h^o$, and any $T \in \mathscr{T}_h$, such that $E \in \partial T$ we set $\boldsymbol{n}_{E,T}$ to be the unit outward (with respect to T) normal vector to E. With every face $E \in \mathscr{E}_h^o$ we also associate a unit vector $\boldsymbol{n}_E$ which is orthogonal to E. For the boundary faces, we always set $\boldsymbol{n}_E = \boldsymbol{n}_{E,T}$, where

T is the *unique* element for which we have $E \subset \partial T$. In our setting, for the interior faces, the particular direction of $\boldsymbol{n}_E$ is not important, although it is important that this direction is fixed. For every face $E \in \mathscr{E}_h$, we define $T^+(E)$ and $T^-(E)$ as follows:

$$\begin{aligned} T^+(E) &:= \{T \in \mathscr{T}_h : \ E \subset \partial T, \ \langle \boldsymbol{n}_E, \boldsymbol{n}_{E,T} \rangle > 0\}, \\ T^-(E) &:= \{T \in \mathscr{T}_h : \ E \subset \partial T, \ \langle \boldsymbol{n}_E, \boldsymbol{n}_{E,T} \rangle < 0\}. \end{aligned} \tag{2}$$

It is clear that for every face we have exactly one $T^+(E)$ and for the interior faces we also have exactly one $T^-(E)$. In the following, we will also write $T^\pm$ instead of $T^\pm(E)$. For a given function $\boldsymbol{w} \in [L_2(\Omega)]^2$ and a fixed interior face $E \in \mathscr{E}_h^o$ the average and jump trace operators are defined by $\{\{\boldsymbol{w}\}\} := (\boldsymbol{w}^+ + \boldsymbol{w}^-)/2$, $[[\boldsymbol{w}]] := (\boldsymbol{w}^+ - \boldsymbol{w}^-)$, where $\boldsymbol{w}^+$ and $\boldsymbol{w}^-$ denote the traces of $\boldsymbol{w}$ onto E taken from within the interior of T^+ and T^-, respectively. On boundary faces $E \in \mathscr{E}_h^\partial$, we set $\{\{\boldsymbol{w}\}\} = \boldsymbol{w}$ and $[[\boldsymbol{w}]] = \boldsymbol{w}$. More about nonconforming discretizations and related solvers for linear elasticity problems can be found in [5–7, 9, 10, 12].

2 Interior Penalty Methods

Let us define the space of piecewise smooth functions

$$[H^2(\mathscr{T}_h)]^2 = \left\{ \boldsymbol{u} \in [L^2(\Omega)]^2 : \boldsymbol{u}\big|_T \in [H^2(T)]^2, \quad \forall\, T \in \mathscr{T}_h \right\},$$

and the linear DG space

$$\boldsymbol{V}^{\mathrm{DG}} = [V^{\mathrm{DG}}]^2, \quad V^{\mathrm{DG}} = \{u \in L^2(\Omega) : \ u\big|_T \in \mathbb{P}^1(T), \quad \forall\, T \in \mathscr{T}_h\},$$

where $\mathbb{P}^1(T)$ is the space of linear polynomials on T. The IP methods can be derived using the weighted residual framework see [2–4, 8] for more details. We may rewrite the continuous problem (1) in the form: find $\boldsymbol{u} \in [H^2(\mathscr{T}_h)]^2$ such that

$$\begin{aligned} -\mathrm{div}\boldsymbol{\sigma}(\boldsymbol{u}) &= \boldsymbol{f} && \text{on } T \in \mathscr{T}_h\,, \\ [[\boldsymbol{\sigma}(\boldsymbol{u})\boldsymbol{n}]]_E &= \boldsymbol{0} && \text{on } E \in \mathscr{E}_h^o\,, \\ [[\boldsymbol{u}]]_E &= \boldsymbol{0} && \text{on } E \in \mathscr{E}_h^o\,, \\ [[\boldsymbol{u}]]_E &= \boldsymbol{0} && \text{on } E \in \mathscr{E}_h^D\,, \\ [[\boldsymbol{\sigma}(\boldsymbol{u})\boldsymbol{n} - \boldsymbol{g}]]_E &= \boldsymbol{0} && \text{on } E \in \mathscr{E}_h^N\,. \end{aligned} \tag{3}$$

We will introduce a variational formulation of (3), see [8], using the following five operators

$$B_0 : [H^2(\mathscr{T}_h)]^2 \longrightarrow [L^2(\mathscr{T}_h)]^2$$

$$B_1, B_2 : [H^2(\mathscr{T}_h)]^2 \longrightarrow [L^2(\mathscr{E}_h^o)]^2, \qquad B_1^\partial, B_2^\partial : [H^2(\mathscr{T}_h)]^2 \longrightarrow [L^2(\mathscr{E}_h^\partial)]^2$$

and weighting each equation in (3) appropriately. Then we will consider the problem: find $\boldsymbol{u} \in [H^2(\mathscr{T}_h)]^2$ such that for all $\boldsymbol{v} \in [H^2(\mathscr{T}_h)]^2$

$$\begin{aligned}(-\mathrm{div}\boldsymbol{\sigma}(\boldsymbol{u}) - \boldsymbol{f}, B_0(\boldsymbol{v}))_{\mathscr{T}_h} + ([[\boldsymbol{\sigma}(\boldsymbol{u})\boldsymbol{n}]], B_2(\boldsymbol{v}))_{\mathscr{E}_h^o} + ([[\boldsymbol{u}]], B_1(\boldsymbol{v}))_{\mathscr{E}_h^o} \\ + ([[\boldsymbol{u}]], B_1^{\partial}(\boldsymbol{v}))_{\mathscr{E}_h^D} + ([[\boldsymbol{\sigma}(\boldsymbol{u})\boldsymbol{n} - \boldsymbol{g}_N]], B_2^{\partial}(\boldsymbol{v}))_{\mathscr{E}_h^N} = \boldsymbol{0}.\end{aligned} \tag{4}$$

To obtain the IP method we set $B_0(\boldsymbol{v}) = \boldsymbol{v}$, $B_2(\boldsymbol{v}) = \{\{\boldsymbol{v}\}\}$ and $B_2^{\partial}(\boldsymbol{v}) = \boldsymbol{v}$. Then from (4) we get

$$\begin{aligned}(-\mathrm{div}\boldsymbol{\sigma}(\boldsymbol{u}), \boldsymbol{v})_{\mathscr{T}_h} + ([[\boldsymbol{\sigma}(\boldsymbol{u})\boldsymbol{n}]], \{\{\boldsymbol{v}\}\})_{\mathscr{E}_h^o \cup \mathscr{E}_h^D} + ([[\boldsymbol{u}]], B_1(\boldsymbol{v}))_{\mathscr{E}_h^o \cup \mathscr{E}_h^D} \\ = (\boldsymbol{f}, \boldsymbol{v})_{\mathscr{T}_h} + (\boldsymbol{g}, \boldsymbol{v})_{\mathscr{E}_h^N}.\end{aligned} \tag{5}$$

Defining

$$\mathscr{F}(\boldsymbol{v}) = (\boldsymbol{f}, \boldsymbol{v})_{\mathscr{T}_h} + ([[\boldsymbol{g}]], \mathscr{B}_1^{\partial}(\boldsymbol{v}))_{\mathscr{E}_h^D} + (\boldsymbol{g}_N, \boldsymbol{v})_{\mathscr{E}_h^N},$$

and integrating by parts the first term on the left side of (5) then leads to

$$(\boldsymbol{\sigma}(\boldsymbol{u}) : \boldsymbol{\varepsilon}(\boldsymbol{v}))_{\mathscr{T}_h} - (\{\{\boldsymbol{\sigma}(\boldsymbol{u})\boldsymbol{n}\}\}, [[\boldsymbol{v}]])_{\mathscr{E}_h^o \cup \mathscr{E}_h^D} + ([[\boldsymbol{u}]], B_1(\boldsymbol{v}))_{\mathscr{E}_h^o \cup \mathscr{E}_h^D} = \mathscr{F}(\boldsymbol{v}).$$

For a fixed edge $E \in \mathscr{E}_h^o \cup \mathscr{E}_h^D$ the operator $B_1(\boldsymbol{v})$ is defined by

$$B_1(\boldsymbol{v}) := -\{\{\boldsymbol{\sigma}(\boldsymbol{v})\boldsymbol{n}\}\} + \alpha_0\beta_0\mathscr{P}_E^0[[\boldsymbol{v}]] + \alpha_1\beta_1[[\boldsymbol{v}]],$$

where, following [11], the parameters β_0 and β_1 are defined as average values depending on the Lamé constants λ and μ:

$$\beta_0 := \{\{3\lambda + 2\mu\}\}, \qquad \beta_1 := \{\{2\mu\}\}.$$

The weak formulation of the linear elasticity problem reads as follows: find $\boldsymbol{u} \in [H^2(\mathscr{T}_h)]^2$ such that

$$\mathscr{A}(\boldsymbol{u}, \boldsymbol{v}) = \mathscr{F}(\boldsymbol{v}), \qquad \forall\ \boldsymbol{v} \in [H^2(\mathscr{T}_h)]^2, \tag{6}$$

where the bilinear form $\mathscr{A}(\cdot, \cdot)$ is given by

$$\mathscr{A}(\boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{\sigma}(\boldsymbol{u}) : \boldsymbol{\varepsilon}(\boldsymbol{v}))_{\mathscr{T}_h} - (\{\{\boldsymbol{\sigma}(\boldsymbol{u})\boldsymbol{n}\}\}, [[\boldsymbol{v}]])_{\mathscr{E}_h} - ([[\boldsymbol{u}]], \{\{\boldsymbol{\sigma}(\boldsymbol{v})\boldsymbol{n}\}\})_{\mathscr{E}_h} + a_j([[\boldsymbol{u}]], [[\boldsymbol{v}]]),$$

and the penalty term $a_j([[\boldsymbol{u}]], [[\boldsymbol{v}]])$ is of the form

$$a_j([[\boldsymbol{u}]], [[\boldsymbol{v}]]) := \alpha_0 \sum_{E \in \mathscr{E}_h} \beta_0 \int_E \langle h_E^{-1}[[\boldsymbol{u}]], \mathscr{P}_E^0[[\boldsymbol{v}]]\rangle + \alpha_1 \sum_{E \in \mathscr{E}_h} \beta_1 \int_E \langle h_E^{-1}[[\boldsymbol{u}]], [[\boldsymbol{v}]]\rangle.$$

Here for a given face E, the operator $\mathscr{P}_E^0 : L^2(E) \mapsto \mathbb{P}^0(E)$ denotes the L^2-projection onto the constant function on E defined by

$$\mathscr{P}_E^0 \boldsymbol{v} = \frac{1}{|E|} \int_E \boldsymbol{v}, \quad \text{for all} \quad \boldsymbol{v} \in [L^2(E)]^2.$$

The parameters α_0 and α_1 are chosen so that the resulting SIPG discretization is consistent and stable, cf. [11].

Finally, to obtain the discrete formulation, we replace $[H^2(\mathscr{T}_h)]^2$ in (6) by $\boldsymbol{V}^{\mathrm{DG}}$, and hence get the discrete problem: find $\boldsymbol{u}_h \in \boldsymbol{V}^{\mathrm{DG}}$ such that

$$\mathscr{A}(\boldsymbol{u}_h, \boldsymbol{v}) = \mathscr{F}(\boldsymbol{v}), \qquad \forall\ \boldsymbol{v} \in \boldsymbol{V}^{\mathrm{DG}}.$$

3 Preconditioning

Let us introduce the classical Crouzeix-Raviart finite element space

$$V^{\mathrm{CR}} = \left\{ v \in L^2(\Omega) \ : \ v_{|T} \ \in \mathbb{P}^1(T), \ \forall T \in \mathscr{T}_h \text{ and } \mathscr{P}_E^0[[v]] = 0, \ \forall\ E \in \mathscr{E}_h^o \right\}.$$

The corresponding space of vector valued functions is

$$\boldsymbol{V}^{\mathrm{CR}} := [V^{\mathrm{CR}}]^2. \tag{7}$$

Following [2] we introduce also the space

$$\mathscr{Z} = \left\{ z \in L^2(\Omega) \ : \ z_{|T} \ \in \ \mathbb{P}^1(T) \ \forall T \in \mathscr{T}_h \text{ and } \mathscr{P}_E^0\{\{z\}\} = 0 \ \forall\ E \in \mathscr{E}_h^o \right\}.$$

The corresponding space of vector valued functions is

$$\boldsymbol{\mathscr{Z}} = [\mathscr{Z}]^2. \tag{8}$$

To describe the basis functions associated with the spaces (7) and (8), let $\varphi_{E,T}$ denote the canonical scalar Crouzeix-Raviart basis function on T, dual to the degree of freedom at the mass center of the face E, and extended as zero outside T. For $E \in \partial T$, $E' \in \partial T$, the function $\varphi_{E,T}(m_{E'}) = \delta_{E,E'}$. Moreover, we have $\varphi_{E,T} \in \mathbb{P}^1(T)$, and $\varphi_{E,T}(x) = 0$ for all $x \notin T$. Observe that any function $\boldsymbol{u} \in \boldsymbol{V}^{\mathrm{DG}}$ can be represented as

$$\begin{aligned} \boldsymbol{u}(x) &= \sum_{T \in \mathscr{T}_h} \sum_{E \in \partial T} \boldsymbol{u}_T(m_E) \varphi_{E,T}(x) \\ &= \sum_{E \in \mathscr{E}_h} \boldsymbol{u}^+(m_E) \varphi_E^+(x) + \sum_{E \in \mathscr{E}_h^o} \boldsymbol{u}^-(m_E) \varphi_E^-(x), \end{aligned} \tag{9}$$

where in the last identity we changed the order of summation and used the notation $\varphi_E^{\pm}(x) := \varphi_{E,T^{\pm}}(x)$ together with

$$\boldsymbol{u}^{\pm}(m_E) := \boldsymbol{u}_{T^{\pm}}(m_E) = \frac{1}{|E|}\int_E \boldsymbol{u}^{\pm} ds, \qquad \forall\, E \in \mathscr{E}_h^o,\ :\ E = \partial T^+ \cap \partial T^-,$$

$$\boldsymbol{u}(m_E) := \boldsymbol{u}_T(m_E) = \frac{1}{|E|}\int_E \boldsymbol{u}_T ds, \qquad \forall\, E \in \mathscr{E}_h^{\partial},\ \text{such that } E = \partial T \cap \partial\Omega.$$

We recall the definitions of $T^+(E)$ and $T^-(E)$ (see Eq. (2)) and set

$$\varphi_E^{CR} = \varphi_{E,T^+(E)} + \varphi_{E,T^-(E)},\ \forall\, E \in \mathscr{E}_h^o, \quad \varphi_E^{CR} = \varphi_{E,T^+(E)},\ \forall\, E \in \mathscr{E}_h^N,$$

and

$$\psi_E^z = \frac{\varphi_{E,T^+(E)} - \varphi_{E,T^-(E)}}{2},\ \forall\, E \in \mathscr{E}_h^o, \qquad \psi_E^z = \varphi_{E,T^+(E)},\ \forall\, E \in \mathscr{E}_h^D.$$

Clearly, $\{\varphi_E^{CR}\}_{E\in\mathscr{E}_h^o\cup\mathscr{E}_h^N}$ are linearly independent, and $\{\boldsymbol{\psi}_{E,T}^z\}_{E\in\mathscr{E}_h^o\cup\mathscr{E}_h^D}$ are linearly independent. A simple calculation then shows that

$$\boldsymbol{V}^{\mathrm{CR}} = \operatorname{span}\left\{\{\varphi_E^{CR}\boldsymbol{e}_k\}_{k=1}^d\right\}_{E\in\mathscr{E}_h^o\cup\mathscr{E}_h^N}, \quad \mathscr{Z} = \operatorname{span}\left\{\{\psi_E^z\boldsymbol{e}_k\}_{k=1}^d\right\}_{E\in\mathscr{E}_h^o\cup\mathscr{E}_h^D}.$$

Here $\boldsymbol{e}_k, k = 1,\dots,d$ is the k-th canonical basis vector in $\mathbb{R}^d$. Hence by performing a change of basis in (9), we have obtained a "natural" splitting

$$\boldsymbol{V}^{\mathrm{DG}} = \boldsymbol{V}^{\mathrm{CR}} \oplus \mathscr{Z},$$

where the set $\{\boldsymbol{\psi}_{E,T}^z\}_{E\in\mathscr{E}_h^o\cup\mathscr{E}_h^D} \cup \{\varphi_E^{CR}\}_{E\in\mathscr{E}_h^o\cup\mathscr{E}_h^N}$ provides a natural basis for the linear DG space. Then we have for any $\boldsymbol{u},\boldsymbol{w} \in \boldsymbol{V}^{\mathrm{DG}}$ we can write $\boldsymbol{u} = \boldsymbol{z} + \boldsymbol{v}$, and $\boldsymbol{w} = \boldsymbol{\psi} + \boldsymbol{\varphi}$, where $\boldsymbol{z},\boldsymbol{\psi} \in \mathscr{Z}$ and $\boldsymbol{v},\boldsymbol{\varphi} \in \boldsymbol{V}^{\mathrm{CR}}$, such that the bilinear form becomes $\mathscr{A}(\boldsymbol{u},\boldsymbol{w}) = \mathscr{A}((\boldsymbol{z},\boldsymbol{v}),(\boldsymbol{\psi},\boldsymbol{\phi}))$. The approximation of $\mathscr{A}(\cdot,\cdot)$ is given by the following block-diagonal preconditioner (see [1,2])

$$\mathscr{B}((\boldsymbol{z},\boldsymbol{v}),(\boldsymbol{\psi},\boldsymbol{\phi})) := \mathscr{A}(\boldsymbol{z},\boldsymbol{\psi}) + \mathscr{A}(\boldsymbol{v},\boldsymbol{\phi}). \tag{10}$$

For the case of homogeneous materials it is shown in [2] that the following inequality holds for any $\boldsymbol{z} \in \mathscr{Z}$ and any $\boldsymbol{v} \in \boldsymbol{V}^{\mathrm{CR}}$

$$\mathscr{A}(\boldsymbol{z},\boldsymbol{v})^2 \le \gamma^2 \mathscr{A}(\boldsymbol{z},\boldsymbol{z})\mathscr{A}(\boldsymbol{v},\boldsymbol{v}),$$

where the constant $\gamma < 1$ is uniformly bounded. This is a basis for the condition number estimate

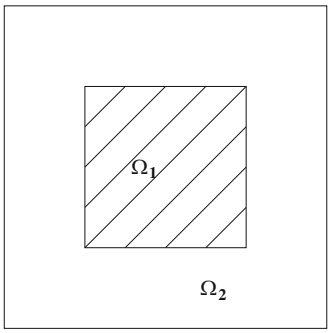


Fig. 1 Computational domain $\Omega = \Omega_1 \cup \Omega_2$

Table 1 Tabulated values of $\kappa(B^{-1}A)$ for different levels of refinement and varying the Poisson ratio ν_2

	$\nu_2 = 0.3$	$\nu_2 = 0.4$	$\nu_2 = 0.49$	$\nu_2 = 0.499$	$\nu_2 = 0.49999$
$\ell = 1$	1.5391	1.3062	1.2685	1.2903	1.2928
$\ell = 2$	1.5457	1.3096	1.2684	1.2900	1.2925
$\ell = 3$	1.5491	1.3115	1.2684	1.2899	1.2924

$$\kappa(B^{-1}A) \leq \frac{1+\gamma}{1-\gamma}, \tag{11}$$

where A and B are the matrices related to the bilinear forms $\mathscr{A}$ and $\mathscr{B}$. Note that the estimate (11) still holds in presence of discontinuous model parameters. In the next section we present the results of two numerical experiments that confirm the robustness of the preconditioner.

4 Numerical Tests

We consider the model problem (1) with mixed boundary conditions on unit square $\Omega = \Omega_1 \cup \Omega_2$ (see Fig. 1). Here

$$\lambda_i = \frac{E_i \nu_i}{(1+\nu_i)(1-2\nu_i)} \quad \text{and} \quad \mu_i = \frac{E_i}{2(1+\nu_i)}$$

are the Lamé parameters in Ω_i, where E_i and ν_i denote the the Young's modulus and the Poisson ratio, respectively. The initial triangulation (level 1) consists of 32 triangles. Each refinement level $\ell+1$ is obtained by subdividing each of the triangles from level (ℓ) into four congruent triangles. The values of the spectral condition numbers have been computed using MATLAB.

Table 2 Condition number of the matrix A_{zz}

	$\nu_2 = 0.3$	$\nu_2 = 0.4$	$\nu_2 = 0.49$	$\nu_2 = 0.499$	$\nu_2 = 0.49999$
$\ell = 1$	4.5433	4.3220	4.0442	4.0047	4.0000
$\ell = 2$	4.5526	4.3250	4.0442	4.0047	4.0000
$\ell = 3$	4.5548	4.3255	4.0442	4.0047	4.0000

Table 3 Condition number $\kappa(B^{-1}A)$ for different levels of refinement and varying Young's modulus E_2

	$E_2 = 1$	$E_2 = 10$	$E_2 = 10^2$	$E_2 = 10^3$	$E_2 = 10^4$
$\ell = 1$	1.2928	1.0913	1.0282	1.0081	1.0292
$\ell = 2$	1.2925	1.0915	1.0283	1.0081	1.0292
$\ell = 3$	1.2924	1.0915	1.0283	1.0081	1.0292

The matrix A_{zz} corresponding to the restriction of $\mathscr{A}(\cdot,\cdot)$ to $\mathscr{Z}$ is well-conditioned and thus the related sub-problem can be solved efficiently using the conjugate gradient method, see Table 2.

The numerical values reported in Tables 1 and 3 confirm the uniform bound of the condition number for problems with jumps in Young's modulus and in the Poisson ratio. Here $E_1 = 1$, $\nu_1 = 0.3$, and $\nu_2 = 0.49999$.

Acknowledgements The authors gratefully acknowledge the support by the Austrian Science Foundation, FWF Project P22989-N18, and also the Bulgarian NSF Grants DO 02-338/08, DMU 03-62/11.

References

1. B. Ayuso, I. Georgiev, J. Kraus, and L. Zikatanov, A simple preconditioner for the SIPG discretization of linear elasticity equations. In I. Dimov, S. Dimova, and N. Kolkovska, editors, *Numerical Methods and Applications*, volume 6046 of *Lecture Notes in Computer Science*, pages 353–360. Springer, Heidelberg, 2011.
2. B. Ayuso, I. Georgiev, J. Kraus, and L. Zikatanov, A Subspace correction method for discontinuous Galerkin discretizations of linear elasticity equations. *arXiv:1110.5743v2*, 2011.
3. B. Ayuso, M. Holst, Y. Zhu and L. Zikatanov, Multilevel preconditioners for discontinuous Galerkin approximation of elliptic problems with jump coefficients. *arXiv:1012/1287vl*, 2010.
4. Blanca Ayuso de Dios and Ludmil Zikatanov. Uniformly convergent iterative methods for discontinuous Galerkin discretizations. *J. Sci. Comput.*, 40(1–3):4–36, 2009.
5. R. Blaheta, S. Margenov, and M. Neytcheva, Aggregation-based multilevel preconditioning of non-conforming FEM elasticity problems. In J. Dongarra, K. Madsen, and J. Wasniewski, editors, *Applied Parallel Computing*, volume 3732 of *Lecture Notes in Computer Science*, pages 847–856. Springer, Berlin, Heidelberg, 2006.
6. S. Brenner and L. Scott The mathematical theory of finite element methods. Texts in applied mathematics. vol. 15, Springer-Verlag, 1994.
7. Susanne C. Brenner and Li-Yeng Sung. Linear finite element methods for planar linear elasticity. *Math. Comp.*, 59(200):321–338, 1992.

8. F. Brezzi, B. Cockburn, L. D. Marini, and E. Süli. Stabilization mechanisms in discontinuous Galerkin finite element methods. *Comput. Methods Appl. Mech. Engrg.*, 195(25–28):3293–3310, 2006.
9. Richard S. Falk. Nonconforming finite element methods for the equations of linear elasticity. *Math. Comp.*, 57(196):529–550, 1991.
10. I. Georgiev, J. K. Kraus, and S. Margenov Multilevel preconditioning of Crouzeix-Raviart 3D pure displacement elasticity problems. In I. Lirkov, S. Margenov, and J. Wasniewski, editors, *Large Scale Scientific Computing*, volume 5910 of *Lecture Notes in Computer Science*, pages 100–107. Springer, Berlin, Heidelberg, 2010.
11. Peter Hansbo and Mats G. Larson. Discontinuous Galerkin and the Crouzeix-Raviart element: application to elasticity. *M2AN Math. Model. Numer. Anal.*, 37(1):63–72, 2003.
12. J. K. Kraus and S. Margenov. *Robust Algebraic Multilevel Methods and Algorithms*. Radon Series on Computational and Applied Mathematics 5. de Gruyter, October 2009.

A Robust Preconditioner for Distributed Optimal Control for Stokes Flow with Control Constraints

M. Kollmann and W. Zulehner

Abstract This work is devoted to the construction and analysis of robust solution techniques for the distributed optimal control problem for the Stokes equations with inequality constraints on the control. There the first order system of necessary and sufficient optimality conditions is nonlinear. A primal-dual active set method is applied in order to linearize the system. In every step a linear saddle point system has to be solved. For this system, we analyze a block-diagonal preconditioner that is robust with respect to the discretization parameter as well as the active set.

1 Introduction

Flow control plays an important role in various fields of applications, see, e.g., [6]. We focus on control by a force term distributed over the whole domain occupied by the fluid, which is of particular interest in relation to electrically conducting fluids that can be influenced by magnetic fields. In this paper we consider velocity tracking problems with stationary and highly viscous flows of incompressible media that are modeled by the Stokes equations. The main focus in these problems is to steer the velocity to a desired state (target velocity) by controlling it by some force, which is typically restricted by inequality constraints. The corresponding nonlinear optimality system can be solved by a primal-dual active set method, which is equivalent to a semi-smooth Newton method (cf. [8]). The resulting linear system

M. Kollmann (✉)
Doctoral Program Computational Mathematics, Johannes Kepler University, Altenbergerstr. 69, A-4040 Linz, Austria
e-mail: markus.kollmann@dk-compmath.jku.at

W. Zulehner
Institute of Computational Mathematics, Johannes Kepler University, Altenbergerstr. 69, A-4040 Linz, Austria
e-mail: zulehner@numa.uni-linz.ac.at

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_81,

in each Newton step is a parameter dependent saddle point problem. In this paper we discuss the preconditioned MinRes method for solving these linear problems robustly with respect to the discretization parameter and the involved active set.

A similar approach is presented in [7] for a distributed optimal control of elliptic equations with various types of inequality constraints, where a preconditioner is constructed based on the mapping properties of the involved operators in Sobolev spaces equipped with the standard norms.

Our construction of a preconditioner and the analysis are based on a result in [15], where a block-diagonal preconditioner for the distributed optimal control problem for the Stokes equations without constraints is presented, which is robust with respect to the discretization parameter and the model parameter. This preconditioner is based on the mapping properties of the involved operators in Sobolev spaces equipped with non-standard norms. Here we use this preconditioner also in the case of constrained optimization and show its robustness with respect to the discretization parameter and the involved active set.

The paper is organized as follows: In Sect. 2 we introduce the velocity tracking problem, discretize it by a finite element method and derive the resulting linearized system. Section 3 deals with the analysis of our block-diagonal preconditioner used in a MinRes method for the linear system. In Sect. 4 we present some numerical experiments. The paper ends with a few concluding remarks.

2 The Optimal Control Problem

As a model problem, we consider the following velocity tracking problem for Stokes flow with distributed control: Find the velocity $\mathbf{u} \in H_0^1(\Omega)^d$, the pressure $p \in L_0^2(\Omega) := \{q \in L^2(\Omega) : \int_\Omega q \, dx = 0\}$, and the force $\mathbf{f} \in L^2(\Omega)^d$ that minimizes the cost functional

$$J(\mathbf{u}, \mathbf{f}) = \frac{1}{2} ||\mathbf{u} - \mathbf{u_d}||^2_{L^2(\Omega)} + \frac{\alpha}{2} ||\mathbf{f}||^2_{L^2(\Omega)} , \tag{1}$$

subject to the state equations

$$\begin{aligned} -\Delta \mathbf{u} + \nabla p &= \mathbf{f} \quad \text{in } \Omega , \quad \operatorname{div} \mathbf{u} = 0 \quad \text{in } \Omega , \\ \mathbf{u} &= \mathbf{0} \quad \text{on } \Gamma , \quad \mathbf{f_a} \le \mathbf{f} \le \mathbf{f_b} \quad \text{a.e. in } \Omega . \end{aligned}$$

Here Ω is an open and bounded domain in $\mathbb{R}^d$ $(d \in \{1, 2, 3\})$ with Lipschitz-continuous boundary Γ, $\mathbf{u}_d \in L^2(\Omega)^d$ is the desired velocity, $\alpha > 0$ is a cost parameter and $\mathbf{f_a}$, $\mathbf{f_b} \in L^2(\Omega)^d$ are the lower and upper bounds for the control variable $\mathbf{f}$, respectively.

There are two possibilities in order to solve this optimal control problem: the first-discretize-then-optimize strategy and the first-optimize-then-discretize strategy. In the situation considered here, both of them result in the same nonlinear

system of equations (see [7, 9] for the elliptic case). In this paper we consider the first-discretize-then-optimize strategy. As an example of a discretization method we discuss the finite element method using the Taylor-Hood element on a simplicial subdivision of Ω consisting of continuous and piecewise quadratic functions for the velocity and continuous and piecewise linear functions for the pressure. Since the optimality conditions in the continuous setting yield a representation of the force in terms of the Lagrange multiplier for the velocity and the Lagrange multiplier for the control constraints (see [7] for the elliptic case), it is reasonable to take the same finite element functions for the discretization of the force as for the velocity.

The discrete counterpart of (1) is:

$$\text{Minimize} \quad \frac{1}{2}(\mathbf{u_h} - \mathbf{u_{d_h}})^T \mathbf{M}(\mathbf{u_h} - \mathbf{u_{d_h}}) + \frac{\alpha}{2}\mathbf{f_h}^T \mathbf{M}\mathbf{f_h} \,, \tag{2}$$

subject to the state equations

$$\begin{aligned} \mathbf{K}\mathbf{u_h} - \mathbf{D}^T\mathbf{p_h} &= \mathbf{M}\mathbf{f_h} \,, \\ -\mathbf{D}\mathbf{u_h} &= \mathbf{0} \,, \\ \mathbf{f_{a_h}} \le \mathbf{f_h} &\le \mathbf{f_{b_h}} \,, \end{aligned}$$

where $\mathbf{M}$ denotes the mass matrix representing the $L^2(\Omega)^d$ scalar product, $\mathbf{K}$ denotes the stiffness matrix representing the vector Laplace operator on the finite element space, $\mathbf{D}$ denotes the divergence matrix representing the divergence operator on the involved finite element spaces and $\mathbf{u_h}$, $\mathbf{f_h}$ and $\mathbf{p_h}$ are the coordinate vectors of $\mathbf{u}$, $\mathbf{f}$ and p w.r.t. the nodal basis, respectively.

The first order system of necessary and sufficient optimality conditions of (2) can be expressed as follows:

$$\left.\begin{aligned} \mathbf{M}\mathbf{u_h} + \mathbf{K}\hat{\mathbf{u}}_\mathbf{h} - \mathbf{D}^T\hat{\mathbf{p}}_\mathbf{h} &= \mathbf{M}\mathbf{u_{d_h}} \,, \\ -\mathbf{D}\hat{\mathbf{u}}_\mathbf{h} &= \mathbf{0} \,, \\ \alpha\mathbf{M}\mathbf{f_h} - \mathbf{M}\hat{\mathbf{u}}_\mathbf{h} + \mathbf{z_h} &= \mathbf{0} \,, \\ \mathbf{K}\mathbf{u_h} - \mathbf{D}^T\mathbf{p_h} - \mathbf{M}\mathbf{f_h} &= \mathbf{0} \,, \\ -\mathbf{D}\mathbf{u_h} &= \mathbf{0} \,, \\ \mathbf{z_h} - \max\left\{\mathbf{0}, \mathbf{z_h} + c(\mathbf{f_h} - \mathbf{f_{b_h}})\right\} - \min\left\{\mathbf{0}, \mathbf{z_h} - c(\mathbf{f_{a_h}} - \mathbf{f_h})\right\} &= \mathbf{0} \,, \end{aligned}\right\} \tag{3}$$

for any $c > 0$ with Lagrange multipliers $\hat{\mathbf{u}}_\mathbf{h}$, $\hat{\mathbf{p}}_\mathbf{h}$ and $\mathbf{z_h}$. These multipliers can be interpreted as sensitivities of the cost functional with respect to variations in the constraints.

In order to solve this system, we propose a primal-dual active set method as introduced in [1]. The resulting system $\mathscr{K}x = b$, which has to be solved in each Newton step, reads (after eliminating $\mathbf{f_h}$ and $\mathbf{z_h}$):

$$\begin{pmatrix} A & B^T \\ B & -C \end{pmatrix} \begin{pmatrix} w_h \\ \hat{w}_h \end{pmatrix} = \begin{pmatrix} e_h \\ g_h \end{pmatrix} , \tag{4}$$

with

$$A = \begin{pmatrix} \mathbf{M} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} , \quad B = \begin{pmatrix} \mathbf{K} & -\mathbf{D}^T \\ -\mathbf{D} & \mathbf{0} \end{pmatrix} = B^T , \quad C = \begin{pmatrix} \alpha^{-1}\mathbf{M}_{\mathbf{C}_{\mathscr{A}}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} ,$$

$$e_h = \begin{pmatrix} \mathbf{M}\mathbf{u}_{\mathbf{d_h}} \\ \mathbf{0} \end{pmatrix} , \quad g_h = \begin{pmatrix} \mathbf{g}_{\mathbf{h_1}} \\ \mathbf{0} \end{pmatrix} , \quad w_h = \begin{pmatrix} \mathbf{u_h} \\ \mathbf{p_h} \end{pmatrix} , \quad \hat{w}_h = \begin{pmatrix} \hat{\mathbf{u}}_\mathbf{h} \\ \hat{\mathbf{p}}_\mathbf{h} \end{pmatrix} ,$$

where

$$\mathbf{M}_{\mathbf{C}_{\mathscr{A}}} = \mathbf{M} - \mathbf{P}_{\mathscr{A}}{}^T \left(\mathbf{P}_{\mathscr{A}} \mathbf{M}^{-1} \mathbf{P}_{\mathscr{A}}{}^T \right)^{-1} \mathbf{P}_{\mathscr{A}} ,$$

$$\mathbf{g}_{\mathbf{h_1}} = \mathbf{P}_{\mathscr{A}}{}^T \left(\mathbf{P}_{\mathscr{A}} \mathbf{M}^{-1} \mathbf{P}_{\mathscr{A}}{}^T \right)^{-1} \left(\mathbf{P}_{\mathscr{A}+} \mathbf{f}_{\mathbf{b_h}} + \mathbf{P}_{\mathscr{A}-} \mathbf{f}_{\mathbf{a_h}} \right) ,$$

and $\mathbf{P}_{\mathscr{A}+}$, $\mathbf{P}_{\mathscr{A}-}$ and $\mathbf{P}_{\mathscr{A}}$ are projection matrices corresponding to the active sets $\mathscr{A}^+$, $\mathscr{A}^-$ and $\mathscr{A} = \mathscr{A}^+ \cup \mathscr{A}^-$, respectively.

The system matrix $\mathscr{K}$ is symmetric and indefinite. For solving the corresponding linear system we propose a MinRes method, see, e.g., [14]. Without preconditioning the convergence rate would deteriorate with respect to the discretization parameter h and the cost parameter α. Therefore, preconditioning is an important issue.

3 A Robust Preconditioning Technique

In this section, we discuss a preconditioning strategy for the saddle point system (4). Due to the symmetry and coercivity properties of the underlying operators the blocks fulfill the following properties: $\mathbf{K} = \mathbf{K}^T > 0$, $\mathbf{M} = \mathbf{M}^T > 0$ and $\mathbf{M}_{\mathbf{C}_{\mathscr{A}}} = \mathbf{M}_{\mathbf{C}_{\mathscr{A}}}{}^T \geq 0$. For our choice of the finite element functions, $\mathbf{D}$ is of full rank.

In [15] a block-diagonal preconditioner is constructed for the distributed optimal control problem of the Stokes equations without constraints on the control. This corresponds to the setting $\mathscr{A} = \emptyset$ in (4). In that case robustness of the block-diagonal preconditioner w.r.t. h and α is shown in [15]. Here we use the same preconditioner for general index sets $\mathscr{A}$ and we will show robustness w.r.t. h and the active set $\mathscr{A}$. Contrary to the case without constraints, robustness w.r.t. α was not observed in numerical experiments for constrained problems.

The proposed preconditioner reads as follows, see [15]:

$$\mathscr{P} = \text{diag}\ (P_1, P_2)\ , \tag{5}$$

where $P_1 = \text{diag}\ \left(\mathbf{P}, \alpha \mathbf{D}\mathbf{P}^{-1}\mathbf{D}^T\right)$ and $P_2 = \alpha^{-1} P_1$ with $\mathbf{P} = \mathbf{M} + \alpha^{1/2}\mathbf{K}$. The next theorem contains the main result of this paper, where we use the following notation: For any symmetric and positive (semi-) definite matrix Q we denote by $\|\cdot\|_Q$ the energy (semi-) norm induced by Q.

Theorem 1. *The system matrix $\mathscr{K}$ of (4) satisfies:*

$$\underline{c}\|x\|_{\mathscr{P}} \leq \sup_{y \neq 0} \frac{y^T \mathscr{K} x}{\|y\|_{\mathscr{P}}} \leq \overline{c}\|x\|_{\mathscr{P}} \quad \forall x\ ,$$

with constants $\underline{c}$, $\overline{c}$ independent of the discretization parameter h and the active set $\mathscr{A}$.

Proof. Due to Theorem 2.6 from [15], it is necessary and sufficient to prove

$$\underline{c}_1 \|w_h\|_{P_1}^2 \leq \|w_h\|_A^2 + \sup_{\hat{r}_h \neq 0} \frac{\left(\hat{r}_h^T B w_h\right)^2}{\left\|\hat{r}_h\right\|_{P_2}^2} \leq \overline{c}_1 \|w_h\|_{P_1}^2\ , \tag{6}$$

$$\underline{c}_2 \left\|\hat{w}_h\right\|_{P_2}^2 \leq \left\|\hat{w}_h\right\|_C^2 + \sup_{r_h \neq 0} \frac{\left(r_h^T B \hat{w}_h\right)^2}{\|r_h\|_{P_1}^2} \leq \overline{c}_2 \left\|\hat{w}_h\right\|_{P_2}^2\ , \tag{7}$$

with $r_h = \begin{pmatrix} \mathbf{v_h} \\ \mathbf{q_h} \end{pmatrix}$, $\hat{r}_h = \begin{pmatrix} \hat{\mathbf{v}}_\mathbf{h} \\ \hat{\mathbf{q}}_\mathbf{h} \end{pmatrix}$ for constants $\underline{c}_1$, $\overline{c}_1$, $\underline{c}_2$ and $\overline{c}_2$ independent of the discretization parameter h and the active set $\mathscr{A}$. For proving (6), we first show

$$\underline{c}_3 \|w_h\|_{P_1} \leq \sup_{\hat{r}_h \neq 0} \frac{\hat{r}_h^T B w_h}{\left\|\hat{r}_h\right\|_{P_1}} \leq \overline{c}_3 \|w_h\|_{P_1}\ , \tag{8}$$

for constants $\underline{c}_3$ and $\overline{c}_3$ independent of the discretization parameter h and the active set $\mathscr{A}$. In order to prove (8) we have to verify the conditions of the Theorem of Brezzi [3]:

The boundedness of the bilinear forms, say a and b, associated with $\mathbf{K}$ and $\mathbf{D}$ is trivial. Using Friedrichs inequality with constant c_F we can show the coercivity of a:

$$a(\mathbf{u_h}, \mathbf{u_h}) = \|\mathbf{u_h}\|_\mathbf{K}^2 \geq \frac{1}{2c_F}\|\mathbf{u_h}\|_\mathbf{M}^2 + \frac{1}{2\sqrt{\alpha}}\sqrt{\alpha}\|\mathbf{u_h}\|_\mathbf{K}^2 \geq \min\left\{\frac{1}{2c_F}, \frac{1}{2\sqrt{\alpha}}\right\}\|\mathbf{u_h}\|_\mathbf{P}^2\ .$$

Since

$$\sup_{\mathbf{v_h}\neq\mathbf{0}} \frac{b(\mathbf{v_h}, \hat{\mathbf{p}}_\mathbf{h})}{\|\mathbf{v_h}\|_\mathbf{P}} = \sup_{\mathbf{v_h}\neq\mathbf{0}} \frac{\hat{\mathbf{p}}_\mathbf{h}{}^T \mathbf{D}\mathbf{v_h}}{\|\mathbf{v_h}\|_\mathbf{P}} = \|\hat{\mathbf{p}}_\mathbf{h}\|_{\mathbf{D}\mathbf{P}^{-1}\mathbf{D}^T} = \frac{1}{\sqrt{\alpha}} \|\hat{\mathbf{p}}_\mathbf{h}\|_{\alpha\mathbf{D}\mathbf{P}^{-1}\mathbf{D}^T} ,$$

the inf-sup condition of b is satisfied. Hence (8) follows.

From (8) and the fact that $P_2 = \alpha^{-1} P_1$ we get

$$\sqrt{\alpha}\underline{c}_3 \|w_h\|_{P_1} \leq \sup_{\hat{r}_h\neq 0} \frac{\hat{r}_h^T B w_h}{\|\hat{r}_h\|_{P_2}} \leq \sqrt{\alpha}\overline{c}_3 \|w_h\|_{P_1} . \tag{9}$$

Furthermore we have

$$0 \leq \|w_h\|_A^2 = \|\mathbf{u_h}\|_\mathbf{M}^2 \leq \|w_h\|_{P_1}^2 . \tag{10}$$

Therefore, combining (9) with (10) yields (6). Equation (7) can be shown analogously. □

As a consequence of Theorem 1 we have:

$$\kappa(\mathscr{P}^{-1}\mathscr{K}) := \|\mathscr{P}^{-1}\mathscr{K}\|_\mathscr{P} \|\mathscr{K}^{-1}\mathscr{P}\|_\mathscr{P} \leq \frac{\overline{c}}{\underline{c}} , \tag{11}$$

i.e., the condition number of the preconditioned system is bounded independently of h and $\mathscr{A}$. Therefore, the number of iterations of the preconditioned MinRes method can be bounded independently of h and $\mathscr{A}$ (see, e.g., [5]).

Remark 1. The result of Theorem 1 can be shown not only on the discrete level but also on the continuous level using the corresponding non-standard norms in $H_0^1(\Omega) \times L_0^2(\Omega)$ for $\mathbf{u}$ and p as well as for the Lagrange multipliers $\hat{\mathbf{u}}$ and $\hat{p}$.

Remark 2. Using the standard norms in $H_0^1(\Omega) \times L_0^2(\Omega)$, as it is done in [7] for the elliptic case, leads to the preconditioner:

$$\mathscr{P}_s = \text{diag}\left(\mathbf{K}, \mathbf{M_p}, \mathbf{K}, \mathbf{M_p}\right) , \tag{12}$$

where $\mathbf{M_p}$ denotes the mass matrix for the pressure element. In this case, one can show a similar result as in Theorem 1.

Remark 3. If we consider the distributed optimal control problem for the Stokes equations with different observation and control domains Ω_1 and Ω_2, we end up with the following linear system:

Table 1 Condition numbers

		α							
k	N	10^{-7}	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	1
4	9,030	>500	70.9	13.92	7.49	8.12	8.68	9.2	9.53
5	36,486	>500	74	14.16	8.25	8.81	9.3	9.72	9.99
6	146,694	>500	79	14.63	8.87	9.46	9.92	10.16	10.34
7	588,294	>500	83	15.21	9.06	9.79	10.25	10.47	10.66

$$\begin{pmatrix} \mathbf{M_1} & \mathbf{0} & \mathbf{K} & -\mathbf{D}^T \\ \mathbf{0} & \mathbf{0} & -\mathbf{D} & \mathbf{0} \\ \mathbf{K} & -\mathbf{D}^T & -\alpha^{-1}\mathbf{M_2} & \mathbf{0} \\ -\mathbf{D} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{u_h} \\ \mathbf{p_h} \\ \hat{\mathbf{u}}_\mathbf{h} \\ \hat{\mathbf{p}}_\mathbf{h} \end{pmatrix} = \begin{pmatrix} \mathbf{M_1 u_{d_h}} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix},$$

where $\mathbf{M_1}$ and $\mathbf{M_2}$ are the mass matrices on Ω_1 and Ω_2, respectively. With the preconditioner $\mathscr{P}$ from above, one can show a similar result as in Theorem 1 with robustness w.r.t. h, Ω_1 and Ω_2.

4 Numerical Experiments

The numerical experiments are carried out on the unit square domain $\Omega = (0,1) \times (0,1) \subset \mathbb{R}^2$. The initial mesh contains four triangles obtained by connecting the two diagonals. The final mesh was constructed by applying k uniform refinement steps to the initial mesh, leading to a meshsize $h = 2^{-k}$. For constructing a practically realizable preconditioner we proceed as follows: First we replace the matrix $\mathbf{D}\left(\mathbf{M} + \alpha^{1/2}\mathbf{K}\right)^{-1}\mathbf{D}^T$ by $\left(\alpha^{1/2}{\mathbf{M_p}}^{-1} + {\mathbf{K_p}}^{-1}\right)^{-1}$ as proposed in [4], where $\mathbf{K_p}$ denotes the stiffness matrix for the pressure element. Then the application of the preconditioner would require the multiplication of a vector from the left by the inverse of the matrices $\mathbf{M} + \alpha^{1/2}\mathbf{K}$, $\mathbf{M_p}$ and $\mathbf{K_p}$. These actions are replaced by one step of a V-cycle iteration for $\mathbf{M} + \alpha^{1/2}\mathbf{K}$ and $\mathbf{K_p}$ and by one step of a symmetric Gauss-Seidel iteration for $\mathbf{M_p}$. The V-cycle is done with one step of a symmetric Gauss-Seidel iteration for the pre-smoothing process and for the post-smoothing process. The resulting realizable preconditioner is spectrally equivalent to the theoretical preconditioner according to the analysis in [2, 10–13].

We demonstrate the efficiency of our solver with two different prescribed active sets.

As a first test case, the active set $\mathscr{A}$ is chosen as the set of all indices of those nodes which lie in the upper half of the computational domain. Table 1 shows the condition number of the preconditioned system matrix with preconditioner $\mathscr{P}$ for various values of h and α, where k denotes the number of refinements, N is the total number of degrees of freedom of the discretized optimality system (4).

Table 2 Condition numbers

		α			
k	N	10^{-12}	10^{-8}	10^{-4}	1
4	9,030	>500	7.34	7.41	9.52
5	36,486	>500	4.95	8.21	9.98
6	146,694	133	6.13	8.88	10.34
7	588,294	16.65	6.71	9.12	10.58

In the second test case, the active set $\mathscr{A}$ is chosen as a randomly distributed set, having the same cardinality as in the first test case. Table 2 shows the condition number of the preconditioned system matrix with preconditioner $\mathscr{P}$.

Additional numerical experiments using the preconditioner $\mathscr{P}_s$ showed that the preconditioner $\mathscr{P}$ has a better performance than the standard one, e.g., while the preconditioner $\mathscr{P}_s$ behaves reasonably only for $\alpha \geq 10^{-2}$, the preconditioner $\mathscr{P}$ behaves reasonably as long as $\alpha \geq 10^{-5}$.

5 Concluding Remarks

In order to develop a robust solver for the linear system (4) we used the block-diagonal preconditioning technique introduced in [15]. The preconditioner constructed there was reused for the control constrained distributed optimal control problem for the Stokes equations and robustness w.r.t. the discretization parameter as well as the active set was shown. Even though the preconditioner is not robust w.r.t. α, the numerical experiments show a good performance of this preconditioner as long as α is not extremely small.

Acknowledgements The research was funded by the Austrian Science Fund (FWF): W1214-N15, project DK12.

References

1. M. Bergounioux, K. Ito, and K. Kunisch. Primal-dual strategy for constrained optimal control problems. *SIAM J. Control Optim.*, 37(4):1176–1194 (electronic), 1999.
2. J. H. Bramble and J. E. Pasciak. Iterative techniques for time dependent Stokes problems. *Comput. Math. Appl.*, 33(1–2):13–30, 1997. Approximation theory and applications.
3. F. Brezzi. On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers. *Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge*, 8(R-2):129–151, 1974.
4. J. Cahouet and J.-P. Chabard. Some fast 3D finite element solvers for the generalized Stokes problem. *Internat. J. Numer. Methods Fluids*, 8(8):869–895, 1988.
5. A. Greenbaum. *Iterative methods for solving linear systems*, volume 17 of *Frontiers in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997.

6. Max D. Gunzburger. *Perspectives in flow control and optimization*, volume 5 of *Advances in Design and Control*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2003.
7. R. Herzog and E. Sachs. Preconditioned conjugate gradient method for optimal control problems with control and state constraints. *SIAM J. Matrix Anal. Appl.*, 31(5):2291–2317, 2010.
8. M. Hintermüller, K. Ito, and K. Kunisch. The primal-dual active set strategy as a semismooth Newton method. *SIAM J. Optim.*, 13(3):865–888 (electronic) (2003), 2002.
9. M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE constraints*, volume 23 of *Mathematical Modelling: Theory and Applications*. Springer, New York, 2009.
10. K.-A. Mardal and R. Winther. Uniform preconditioners for the time dependent Stokes problem. *Numer. Math.*, 98(2):305–327, 2004.
11. K.-A. Mardal and R. Winther. Uniform preconditioners for the time dependent Stokes problem. *Numer. Math.*, 103(1):171–172, 2006.
12. M. A. Olshanskii, J. Peters, and A. Reusken. Uniform preconditioners for a parameter dependent saddle point problem with application to generalized Stokes interface equations. *Numer. Math.*, 105(1):159–191, 2006.
13. M. A. Olshanskii and A. Reusken. On the convergence of a multigrid method for linear reaction-diffusion problems. *Computing*, 65(3):193–202, 2000.
14. C. C. Paige and M. A. Saunders. Solutions of sparse indefinite systems of linear equations. *SIAM J. Numer. Anal.*, 12(4):617–629, 1975.
15. W. Zulehner. Non-standard Norms and Robust Estimates for Saddle Point Problems. *SIAM J. Matrix Anal. Appl.*, 32(2):536–560, 2011.

Computing Inner Eigenvalues of Matrices in Tensor Train Matrix Format

T. Mach

Abstract The computation of eigenvalues is one of the core topics of numerical mathematics. We will discuss an eigenvalue algorithm for the computation of inner eigenvalues of a large, symmetric, and positive definite matrix M based on the preconditioned inverse iteration

$$x_{i+1} = x_i - B^{-1}\left(Mx_i - \mu(x_i)x_i\right),$$

and the folded spectrum method (replace M by $(M-\sigma I)^2$). We assume that M is given in the tensor train matrix format and use the TT-toolbox from I.V. Oseledets (see http://spring.inm.ras.ru/osel/) for the numerical computations. We will present first numerical results and discuss the numerical difficulties.

1 Introduction

Let $M \in \mathbb{R}^{m\times m}$ be a matrix. If the pair (λ, v) fulfills

$$Mv = \lambda v, \tag{1}$$

then λ is called an eigenvalue and v an eigenvector of M. They are computed in several applications like structural and vibrational analysis or quantum molecular dynamics. If M is small and dense, then this problem is almost solved. There are good algorithms for the computation of all eigenvalues, for instance the implicit, multi-shift QR algorithm with aggressive early deflation [2]. For large sparse

T. Mach (✉)
Max Planck Institute for Dynamics of Complex Technical Systems, Sandtorstr. 1, 39106 Magdeburg, Germany
e-mail: thomas.mach@googlemail.com

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_82,

matrices the Jacobi-Davidson algorithm [17] or the preconditioned inverse iteration [4], PINVIT for short, can be used to compute some eigenvalues.

For large matrices the dense approach is not feasible, since it requires m^2 storage entries and $\mathcal{O}(m^3)$ flops. Sparse matrices require only $\mathcal{O}(m)$ storage such that large matrices can be handle using sparse matrix arithmetic. But even $\mathcal{O}(m)$ may become too expensive for large m. The tensor-trains, see Sect. 1.1, are one way out, since the storage complexity is in $\mathcal{O}(\log_2 m)$.

The main feature of PINVIT is the matrix dimension independent convergence. This makes preconditioned inverse iteration a preferable method for the computation of the smallest eigenvalues of a matrix given in a compressed storage scheme like the tensor train matrix format. The combination with the folded spectrum method permits also the computation of inner eigenvalues. In [1], this was investigated for the data-sparse hierarchical matrices by the author.

1.1 Tensor Trains

The concept of tensor trains, TT for short, as described in [14] permits the computation of a data-sparse approximation of tensors. Therefore, the tensor $T \in \mathbb{R}^{n^d}$ is approximated by

$$T = \sum_{\alpha_1,\ldots,\alpha_d=1}^{r_1,\ldots,r_d} T_1(i_1,\alpha_1)T_2(\alpha_1,i_2,\alpha_2)\cdots T_{d-1}(\alpha_{d-2},i_{d-1},\alpha_{d-1})T_d(\alpha_{d-1},i_d). \quad (2)$$

The ranks r_i of the summations are called the local ranks of the approximation. The smallest r with $r \geq r_i \ \forall i$ is the local rank of the tensor train. The TT decomposition for $d = 2$ is the low rank factorization

$$\mathbb{R}^{n^2} \ni T = \operatorname{vec}\left(AB^T\right), \quad A, B \in \mathbb{R}^{n\times r}.$$

For the full tensor we have to store n^d entries, but for the approximation in the TT format $(d-2)nr^2 + 2nr$ entries are sufficient. The main advantage is that the storage complexity grows only linearly with d in the TT format.

There are arithmetic operations with tensor trains available in the tensor train toolbox [13] for MATLAB®. We will use here the TT dot product, the addition of two tensor trains, and the rounding of a tensor train, see [12]. These operations require all $\mathcal{O}(dnr^3)$ flops.

1.2 Tensor Train Matrix Format

If we have given a vector $v \in \mathbb{R}^{n^2}$ representing the values of a function over a 2D regular grid, then we can regard v as a vector over the product index set (i_1, i_2), with $i_1 \in I_1$ and $i_2 \in I_2$, where I_1 is the index set for the points in the one direction and

I_2 for the points in the other direction. The generalization of this concept leads to a tensor.

A matrix over the same grid would be described by $((i_1, i_2), (j_1, j_2))$, with (i_1, i_2) as row index and (j_1, j_2) as column index. The generalization leads to a tensor with the structure

$$T(i_1, \ldots, i_d; j_1, \ldots, j_d),$$

where the semicolon separates the row and the column indices. Of course one can use again a tensor train approximation, but now with two indices per tensor carriage:

$$M = \sum_{\alpha_1, \ldots, \alpha_d} M_1(i_1, j_1, \alpha_1) M_2(\alpha_1, i_2, j_2, \alpha_2) \cdots M_d(\alpha_{d-1}, i_d, j_d). \tag{3}$$

This data-sparse matrix format was recently invented by Oseledets, see [12]. A matrix in this format is called a *TTM matrix* or a *matrix in tensor train matrix format*.

If one regards (i_k, j_k) as one long index of length n^2, then the same data structure as in the tensor train case can be used. This immediately permits us to perform additions and rounding with in the TTM matrices. Further the dot product can be used to compute the Frobenius norm of a TTM matrix in $\mathcal{O}(dn^2r^3)$ flops. We will further use the TTM-TT product, see [11].

1.3 Problem Setting

We assume the matrix $M \in \mathbb{R}^{2^d \times 2^d}$ is given in tensor train matrix format. The task is to compute an eigenvector v of M. Since 2^d might be large we are satisfied with an approximation to v in the tensor train format.

We will use preconditioned inverse iteration (PINVIT) to compute the smallest eigenvalue and the corresponding eigenvector of M. A similar approach was already investigated by Lebedeva in [7], but not for TTM matrices. Here we will extend this to the computation of inner eigenvalues by using the folded spectrum method.

In [6] Christine Tobler and Daniel Kressner investigated the usage of LOBPCG for the solution of eigenvalues problems with matrices in $\mathscr{H}$-Tucker format. The LOBPCG method is a variant of PINVIT that uses the optimal vector in the space spanned by the two previous iterates and the preconditioned residual as next iterate.

2 Preconditioned Inverse Iteration

PINVIT can be motivated as an inexact Newton-method for the minimization of the Rayleigh quotient. The Rayleigh quotient $\mu(x)$ for a vector x and a symmetric, positive definite matrix M is defined by

Algorithm 1: Subspace preconditioned inverse iteration

Input: $M \in \mathbb{R}^{m\times m}$, $X_0 \in \mathbb{R}^{m\times s}$ e. g. randomly chosen
Output: $X_p \in \mathbb{R}^{m\times s}$, $\mu \in \mathbb{R}^{s\times s}$, with $\|MX_p - X_p\mu\| \leq \epsilon$
$T^{-1} \approx (M)^{-1}$;
Orthogonalize X_0;
$\mu := X_0^T M X_0$;
$R := \texttt{round}(MX_0 - X_0\mu, \epsilon)$;
$i := 1$;
while $\|R\|_F > \epsilon$ **do**
  $i := i + 1$;
  $X_i := \texttt{round}(X_{i-1} - T^{-1}R, \epsilon)$;
  Orthogonalize X_i;
  $R := \texttt{round}(MX_i - X_i\mu, \epsilon)$ with $\mu := X_i^T M X_i$;
end

$$\mu(x, M) := \mu(x) = \frac{x^T M x}{x^T x}. \tag{4}$$

The global minimum of $\mu(x)$ is reached for $x = v_1$, with $\lambda_1 = \mu(x)$. This means that minimizing the Rayleigh quotient is equal to computing the smallest eigenvalue. Doing this minimization by the following inexact Newton method:

$$x_{i+1} = x_i - B^{-1}\left(Mx_i - \mu(x_i)x_i\right), \tag{5}$$

we get the update equation of preconditioned inverse iteration. The preconditioner B^{-1} for M have to fulfill

$$\left\| I - B^{-1}M \right\|_M \leq c < 1. \tag{6}$$

This method is know for a long time and was, among others, extensively investigated by Knyazev and Neymeyr, see [4, 5, 9, 10]. The main feature of preconditioned inverse iteration is the independence of the convergence from the matrix dimension m.

Now we apply the preconditioned inverse iteration to a matrix in tensor train matrix format. A subspace version of preconditioned inverse iteration is listed in Algorithm 1. First we have to compute the preconditioner. We will use here the Newton-Schulz iteration [16] see Algorithm 2 that is based on the modification of the Newton-Schulz method for TTM matrices in [15]. In the remainder of the Algorithm 1 only TT dot products, TTM-TT products, and additions are used.

3 Computing Inner Eigenvalues by Folded Spectrum Method

Sometimes one is also interested in the inner eigenvalues M. Unfortunately a simple shifting, $M - \mu I$, is not possible, since the preconditioned inverse iteration requires positive definiteness of M. One way out is the use of the so called folded spectrum

Algorithm 2: Newton-Schulz Inversion

Input: $M \in \mathbb{R}^{m \times m}$
Output: $M^{-1} \in \mathbb{R}^{m \times m}$
$Y = M / \|M\|_2$; $X = I / \|M\|_2$;
while $\|Y - I\| > \epsilon_c$ **do**
  $H = \texttt{round}(2I - Y, \epsilon)$; $Y = \texttt{round}(YH, \epsilon)$; $X = \texttt{round}(HX, \epsilon)$;
end
$M^{-1} = X$;

method, [18], which was before also mentioned in [8]. The key observation are the following facts: First, the matrix $M_\mu = (M - \mu I)^2$ is positive definite. If M is symmetric, then M_μ is symmetric, too. Second, an eigenvector v of M to the eigenvalue λ is also an eigenvector of M_μ, since

$$\begin{aligned} M_\mu v &= (M - \mu I)^2 v = M^2 v - 2\mu M v + \mu^2 v \\ &= \lambda^2 v - 2\mu\lambda v + \mu^2 v = (\lambda - \mu)^2 v. \end{aligned}$$

The computation of inner eigenvalues consists of the following steps:

- Choose a shift μ.
- Compute $M_\mu = (M - \mu I)^2$.
- Use PINVIT to compute the smallest eigenpair (v, λ) of M_μ.
- The sought eigenpair is $(v, v^T M v / v^T v)$.

As the numerical experiments in the next section show this procedure can be used to compute some inner eigenvalues of a TTM matrix. But this algorithm have two drawbacks. First the condition number of M_μ is approximately the square of the condition number of M. This means that it is more difficult to invert M_μ. The first consequence is that the Newton-Schulz iteration takes longer. Second, the approximate inverse has larger local ranks and so more storage is necessary. Third, the larger local ranks make the application of the preconditioner more expensive, such that the preconditioning in each step takes also longer. The numerical experiments confirm this.

The second drawback is that the squaring "fold" the spectrum. It may happen that two eigenvalues on different sides of μ become (almost) equal after the squaring. In this case the preconditioned inverse iteration will compute an vector in the invariant subspace spanned by the eigenvectors of both eigenvalues. So one should carefully choose the shift in a way not producing new multiple eigenvalues. Further the computation of the whole invariant subspace lead to good approximations of the sought eigenvalues.

Table 1 Numerical results 2D Laplace, three smallest eigenvalues

n	d	t_{inv} in s	t_{PINVIT} in s	# it	Error
1,024	5	1.038	1.933	18	2.5084 e−07
4,096	6	2.637	4.371	17	1.8998 e−07
16,384	7	5.174	9.132	19	7.3614 e−08
65,536	8	9.233	25.326	21	6.8874 e−08
262,144	9	16.240	37.645	19	1.5501 e−08
1,048,576	10	27.387	48.643	21	2.5858 e−10
4,194,304	11	99.911	146.910	25	1.5207 e−09
16,777,216	12	140.043	154.761	23	9.9111 e−10
67,108,864	13	528.491**	348.101	20	2.8968 e−08
268,435,456	14	1,064.433***	767.721	26	1.5802 e−07
1,073,741,824	15	1,919.606***	2,767.084	53	7.4038 e−07
4,294,967,296	16	3,423.903****	2,796.697	28	6.8776 e−07

Table 2 Numerical results 2D Laplace, shifted ($\mu = 203.3139$), three eigenvalues

n	d	t_{inv} in s	t_{PINVIT} in s	# it	Error
256	4	2.832	2.827	47	8.9370 e−09
1,024	5	15.632	3.431	25	5.0051 e−11
4,096	6	47.027	13.979	33	2.5580 e−12
16,384	7	986.699**	39.519	35	4.8601 e−12
65,536	8	248.160	1,354.844	100	1.4737 e−01

Table 3 Numerical results 3D Laplace, four smallest eigenvalues

n	d	t_{inv} in s	t_{PINVIT} in s	# it	Error
32,768	5	4.045	40.870	30	2.5526 e−07
262,144	6	10.305	129.937	25	2.4926 e−07
2,097,152	7	18.168	473.681	23	1.4520 e−07
16,777,216	8	37.216	2,084.095	24	3.0187 e−08
134,217,728	9	97.761	14,432.496	100	1.9257 e−06
1,073,741,824	10	93.512	4,425.800	21	1.6571 e−08

Table 4 Numerical results 3D Laplace, shifted ($\mu = 230.6195$), six eigenvalues

n	d	t_{inv} in s	t_{PINVIT} in s	# it	Error
64	2	0.289	11.450	100	5.9918 e−06
512	3	8.174	18.930	30	2.5871 e−07
4,096	4	99.393	117.183	67	2.9559 e−12

4 Numerical Experiments

We implement Algorithm 1 using the tensor train toolbox for MATLAB, [13] (source code: http://www.mpi-magdeburg.mpg.de/preprints/2011/1109/). As example we use the Laplace-equation over the unit-square in 2, 3 or 4 dimensions.

Table 5 Numerical results 4D Laplace, five smallest eigenvalues

n	d	t_{inv} in s	t_{PINVIT} in s	# it	Error
65,536	4	1.885	58.279	28	2.8761 e−07
1,048,576	5	5.719	260.385	28	2.4963 e−07
16,777,216	6	15.194	1,028.412	25	1.6518 e−07
268,435,456	7	29.104	5,282.001	35	2.9061 e−09

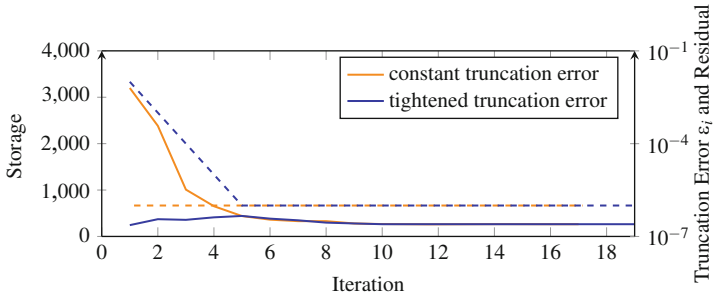


Fig. 1 Memory (*left axis*, *solid line*) used for $X^{(i)}$ depending on the truncation error ϵ_i (*right axis*, *dashed line*); 2D-Laplace example, $d = 6$

We stop the iteration after 100 steps or if the Frobenius norm of the residual drops below 10^{-5}. Tables 1–5 show the results. We choose the preconditioner accuracy $c = 0.2$. Inside the iteration we truncate each tensor trains to a precision of 10^{-6}. If the convergence stagnates we reduce the truncation tolerance. We observe that the local rank of the approximation in the first steps can be very high compared with the local rank in the final approximation. Therefore it makes sense to start with precision 10^{-2} in the first iteration and tighten this by a factor 10 in the first four steps. The effect is shown in Fig. 1. This idea was described in [3] for general truncated iterations.

The tables show the size of the matrix in the first column, d in the second. We have 2^d discretization points in each direction. The third column gives the time for the inversion and the fourth column the time for the preconditioned inverse iteration. This is followed by the number of iterations. The error of the computed eigenvalues, which you found in the last column, is measured by $\|((\lambda_i - \hat{\lambda}_i)/\lambda_i)_{i=1}^{s}\|_\infty$.

One can see that the number of iterations is almost independent from the matrix dimension. The relative error is smaller than 10^{-5}. We further observe that the required CPU time grows slower than the matrix dimension m. This permit the computation of the eigenvalues of large matrices.

The computation for the shifted and squared matrix M_μ is much more expensive than for M. This confirms the expectations from the last section. The large local ranks of the approximation of M_μ^{-1} limit the usage of the method to the computation of inner eigenvalues of comparable small matrices.

5 Conclusions

We have seen that the preconditioned inverse iteration can be used to compute the smallest eigenvalues of a matrix in tensor train matrix format. Further the folded spectrum method makes PINVIT also applicable for the computation of inner eigenvalues, but the numerical examples showed that is much more expensive.

Acknowledgements The author thanks Peter Benner for suggesting to investigate the combination of preconditioned inverse iteration and folded spectrum method also for tensor trains.

References

1. P. Benner and T. Mach. The preconditioned inverse iteration for hierarchical matrices. *Numer. Lin. Alg. Appl.*, 2012. Published online, 17 pages.
2. K. Braman, R. Byers, and R. Mathias. The multi-shift QR-algorithm: Aggressive early deflation. *SIAM J. Matrix Anal. Appl.*, 23(4):948–989, 2002.
3. W. Hackbusch, B.N. Khoromskij, and E. E. Tyrtyshnikov. Approximate iterations for structured matrices. *Numer. Math.*, 109:365–383, 2008.
4. A. V. Knyazev. Preconditioned eigensolvers — an oxymoron? *Electr. Trans. Num. Anal.*, 7:104–123, 1998.
5. A. V. Knyazev and K. Neymeyr. Gradient flow approach to geometric convergence analysis of preconditioned eigensolvers. *SIAM J. Matrix Anal. Appl.*, 31(2):621–628, 2009.
6. D. Kressner and C. Tobler. Preconditioned low-rank methods for high-dimensional elliptic PDE eigenvalue problems. *Comput. Methods Appl. Math.*, 11(3):363–381, 2011.
7. O. S. Lebedeva. Block tensor conjugate gradient-type method for Rayleigh quotient minimization in two-dimensional case. *Comput. Math. Math. Phys.*, 50:749–765, 2010. 10.1134/S0965542510050015.
8. R. B. Morgan. Computing interior eigenvalues of large matrices. *Linear Algebra Appl.*, 154–156:289–309, 1991.
9. K. Neymeyr. A geometric theory for preconditioned inverse iteration I: Extrema of the Rayleigh quotient. *Linear Algebra Appl.*, 322(1–3):61–85, Jan. 2001.
10. K. Neymeyr. *A Hierarchy of Preconditioned Eigensolvers for Elliptic Differential Operators*. Habilitationsschrift, Mathematische Fakultät der Universität Tübingen, Sep. 2001.
11. I. V. Oseledets. Approximation of $2^d \times 2^d$ matrices using tensor decomposition. *SIAM J. Matrix Anal. Appl.*, 31(4):2130–2145, 2010.
12. I. V. Oseledets. Tensor-train decomposition. *SIAM J. Sci. Comput.*, 33(5):2295–2317, 2011.
13. I. V. Oseledets. TT-toolbox 2.2. http://spring.inm.ras.ru/osel/, 2012.
14. I. V. Oseledets, D. V. Savostyanov, and E. E. Tyrtyshnikov. Linear algebra for tensor problems. *Computing*, 85:169–188, 2009.
15. I. V. Oseledets and E. E. Tyrtyshnikov. Breaking the curse of dimensionality or how to use SVD in many dimensions. *SIAM J. Sci. Comput.*, 31(5):3744–3759, 2009.
16. G. Schulz. Iterative Berechung der reziproken Matrix. *ZAMM Z. Angew. Math. Mech.*, 13:57–59, 1933.
17. G. L. G. Sleijpen and H. A. Van der Vorst. A Jacobi-Davidson iteration method for linear eigenvalue problems. *SIAM Rev.*, 42(2):267–293, 2000.
18. L.-W. Wang and A. Zunger. Electronic structure pseudopotential calculations of large (∼1000 atoms) Si quantum dots. *J. Phys. Chem.*, 98(98):2158–2165, 1994.

Two Mathematical Tools to Analyze Metastable Stochastic Processes

T. Leliévre

Abstract We present how entropy estimates and logarithmic Sobolev inequalities on the one hand, and the notion of quasi-stationary distribution on the other hand, are useful tools to analyze metastable overdamped Langevin dynamics, in particular to quantify the degree of metastability. We discuss the interest of these approaches to estimate the efficiency of some classical algorithms used to speed up the sampling, and to evaluate the error introduced by some coarse-graining procedures. This paper is a summary of a plenary talk given by the author at the ENUMATH 2011 conference.

1 Introduction and Motivation

The aim of this paper is to present two mathematical viewpoints on metastability. Roughly speaking, a dynamics is said to be metastable if it spends a lot of time in a region (called a metastable state) before hopping to another region. To be more specific, we will focus in the following on the overdamped Langevin dynamics:

$$dX_t = -\nabla V(X_t)\, dt + \sqrt{2\beta^{-1}} dW_t \tag{1}$$

which is used for example in molecular dynamics to describe the evolution of a molecular system. In this context, the configuration of the system $X_t \in \mathbb{R}^n$ is the coordinates of the particles (think of the atoms of a large molecule), $V : \mathbb{R}^n \to \mathbb{R}$ is the potential energy, which to a configuration $x \in \mathbb{R}^n$ associates its energy $V(x)$, and

T. Leliévre (✉)
CERMICS, Ecole des Ponts Paristech, Université Paris-Est, 6 et 8 avenue Blaise Pascal, 77455, Marne-la-Vallée, France

INRIA Rocquencourt, MicMac project team, Domaine de Voluceau, B.P. 105, 78153, Le Chesnay Cedex, France
e-mail: lelievre@cermics.enpc.fr

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_83,

$\beta^{-1} = k_B T$ is proportional to the temperature (k_B being the Boltzmann constant). The stochastic process W_t is a standard n-dimensional Brownian motion. For such a dynamics, metastability typically originates from two mechanisms. First, in the small temperature regime, the dynamics (1) can be seen as a perturbation of the simple gradient dynamics $\dot{y} = -\nabla V(y)$ for which, from any initial condition, the solution converges to a local minimum of V. Having this in mind, the dynamics (1) is metastable because it takes a lot of time to leave the vicinity of a local minimum before jumping to the neighborhood of another local minimum. This is due to the *energy barriers* which have to be overcome. Such barriers and the zero temperature limit can be analyzed in particular with large deviation techniques [19]. Second, metastability may come from entropic effects. Imagine that the configuration space is made of two boxes linked by a narrow corridor (the potential V is zero on this configuration space, and infinite outside). Then, the dynamics (1) is metastable because it takes a lot of time to find the small corridor to go from one box to the other. Metastability is here due to *entropic barriers*. This can be quantified using the notion of free energy, see [29, 36]. In practice, metastability thus originates from a combination of both energetic and entropic effects, the relative importance of each depending on the system under consideration, and on the temperature. See Fig. 1 for a numerical illustration. Generally speaking, metastability is related to the multimodality of the measure μ, namely the fact that some high probability regions are separated by low probability regions.

An important concept related to metastability is ergodicity. Under adequate assumptions on V, the dynamics (1) can be shown to be ergodic with respect to the canonical measure:

$$\mu(dx) = Z^{-1} \exp(-\beta V(x))\, dx, \tag{2}$$

where $Z = \int_{\mathbb{R}^n} \exp(-\beta V(x))\, dx$ is assumed to be finite. Ergodicity actually refers to two different properties: (i) an average along a trajectory converges in the long-time limit towards an average with respect to μ: for any test function $\varphi : \mathbb{R}^n \to \mathbb{R}$,

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T \varphi(X_t)\, dt = \int_{\mathbb{R}^n} \varphi d\mu, \tag{3}$$

and (ii) the law of the process X_t at time t converges to μ in the long-time limit: for any test function $\varphi : \mathbb{R}^n \to \mathbb{R}$,

$$\lim_{T\to\infty} \mathbb{E}(\varphi(X_T)) = \int_{\mathbb{R}^n} \varphi d\mu. \tag{4}$$

If X_t is metastable, both limits (3) and (4) are typically very difficult to reach, since T should be sufficiently large to visit all the metastable states. From a numerical viewpoint, metastability raises thus sampling issues, both to compute canonical averages (namely averages with respect to μ) and to compute averages over paths

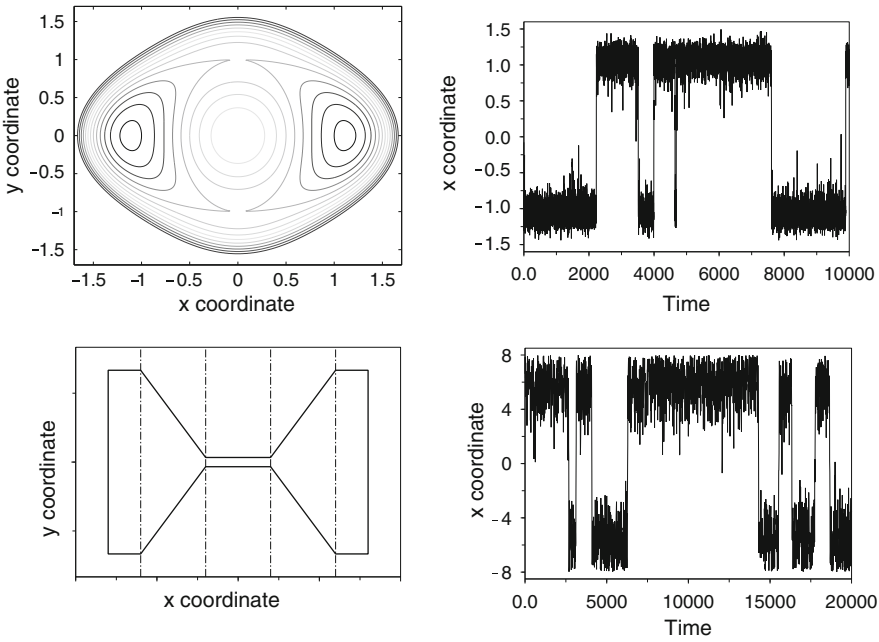


Fig. 1 Above: (*left*) an example of a 2 dimensional potential, with energetic barriers (there are actually two possible saddle points to go from *left* to *right*), with (*right*) the x-component as a function of time of the associated stochastic process solution to (1). Below (*left*) an example of a 2 dimensional potential (which is zero inside the closed *solid-line shape*, and infinite outside) for which an entropic barrier (*right*) is observed

which requires to generate efficiently metastable dynamics, the latter being of course a more complicated task than the former.

In the following, we would like to introduce two mathematical tools to measure the "degree of metastability" of the dynamics (1). In Sect. 2, we discuss the notion of *Logarithmic Sobolev Inequality (LSI)*, which is a way to quantify the ergodic features of a process, and more precisely, how fast the convergence (4) happens. In this context, the slower the convergence, the more metastable the process is. In Sect. 3, we introduce the notion of *quasi-stationary distribution*, and identify the typical time it takes, for a given region of the state space, to reach a quasi equilibrium in this region before leaving it. Metastability in this case is related to the fact that this time is small compared to the typical time it takes to leave the region. In both cases, we will explain how these tools can be used to (i) analyze numerical methods which are used in molecular dynamics to "accelerate" metastable dynamics (see Sects. 2.3 and 3.2) and (ii) obtain coarse-grained descriptions of the metastable dynamics (see Sects. 2.4 and 3.3).

We would like to emphasize the importance of quantifying metastability for practical aspects. Indeed, there exists in the literature many asymptotic analysis in some limiting regimes (zero temperature limit, time-scale separation limit enforced through an explicit small parameter introduced in the dynamics, etc...) where

it is shown how a metastable dynamics converges to some effective Markovian dynamics. In practice, the parameters which are considered to go to zero in the asymptotic regimes may indeed be small, but are certainly not zero. A natural question is to quantify the error introduced by assuming that these parameters are zero, an assumption which is behind many numerical methods, and coarse-graining approaches. This requires in turn quantifying the metastable features of the original dynamics. This is precisely the aim of both approaches presented below.

Let us finally clearly state that the aim of this paper is not to provide proofs of the announced results and sometimes even not to state them very precisely mathematically, but to gather in a new and hopefully enlightening way various recent studies, in particular [28,29,35]. All the statements below can be reformulated as precise mathematical claims, with rigorous proofs, except the discussion in Sect. 3.3 which is more prospective.

Remark 1. There are other techniques to quantify metastability, that we do not review here. We would like to mention in particular spectral approaches [24, 44], potential theoretic approaches [3, 4] and approaches based on drift conditions [26]. Drawing connections between these various techniques is an interesting subject, see for example [7] for connections between approaches based on drift conditions, and functional inequalities such as LSI.

Remark 2. In this paper, we concentrate on the overdamped Langevin dynamics (1) even though this is not the most widely used dynamics in molecular dynamics. All the algorithms we present below generalize to (and are used with) the phase-space Langevin dynamics, which is much more popular. However, generalizing the mathematical approaches outlined below to the Langevin dynamics is not an easy task, due to the lack of ellipticity of the associated infinitesimal generator, see [23, 47] for examples of studies in that direction.

2 Logarithmic Sobolev Inequality

As explained above, we quantify in this section the metastability of (1) by considering the rate of convergence of the limit (4). We thus consider the law at time t of X_t, which has a density $\psi(t,x)$. The probability density function ψ satisfies the Fokker-Planck equation:

$$\partial_t \psi = \mathrm{div}(\nabla V \psi + \beta^{-1} \nabla \psi). \tag{5}$$

Notice that the density of μ with respect to the Lebesgue measure, denoted by $\psi_\infty(x) = Z^{-1} \exp(-\beta V(x))$, is obviously a stationary solution to (5).

2.1 Definition

Let us introduce the notion of logarithmic Sobolev inequality (see for example [1]).

Definition 1. The probability measure μ is said to satisfy a logarithmic Sobolev inequality with constant R (in short LSI(R)) if and only if, for any probability measure ν absolutely continuous with respect to μ,

$$H(\nu|\mu) \leq \frac{1}{2R} I(\nu|\mu)$$

where $H(\nu|\mu) = \int_{\mathbb{R}^n} \ln\left(\frac{d\nu}{d\mu}\right) d\nu$ is the relative entropy of ν with respect to μ, and $I(\nu|\mu) = \int_{\mathbb{R}^n} \left|\nabla \ln\left(\frac{d\nu}{d\mu}\right)\right|^2 d\nu$ is the Fisher information of ν with respect to μ.

When both μ and ν admit densities (respectively ψ and ϕ) with respect to the Lebesgue measure, we shall also use the notation $H(\psi|\phi)$ for $H(\mu|\nu)$. A crucial property is the following:

Proposition 1. *The measure μ with density ψ_∞ defined by (2) satisfies a LSI(R) if and only if, for all probability density functions ψ_0, for all time $t \geq 0$,*

$$H(\psi(t,\cdot)|\psi_\infty) \leq H(\psi_0|\psi_\infty) \exp(-2\beta^{-1} R t) \tag{6}$$

where ψ is the solution to (5) with initial condition $\psi(0,\cdot) = \psi_0$.

This is a simple consequence of the standard computation: if ψ satisfies (5), then

$$\frac{d}{dt} H(\psi(t,\cdot)|\psi_\infty) = -\beta^{-1} I(\psi(t,\cdot)|\psi_\infty).$$

A natural way to quantify metastability if thus to relate it to R:

$$\text{The smaller } R\text{, the more metastable the dynamics (1) is.} \tag{7}$$

Before we proceed, let us make two remarks. First, by combining the classical Bakry-Emery criteria with the perturbation result of Holley and Stroock, the measure μ actually satisfies a LSI under very mild assumptions on V: basically, if V is smooth, and is α-convex at infinity, then a LSI for μ holds. What is more complicated is to get the optimal constant R. Second, in the simple case of a double well potential in dimension 1, it is standard to show that the average time spent by the process X_t in a well before hopping to another one increases exponentially fast as the temperature goes to zero, using for example large deviation techniques. It can also be checked in this simple setting that the constant R scales like $\exp(-\beta\delta V)$ in the limit of large β (small temperature), where δV is the height of the barrier to overcome to leave a given well. This prototypical situation thus shows that the LSI constant indeed allows to quantifying the intuitive definition of metastability we gave in the introduction.

2.2 Metastability Along a Reaction Coordinate

Many algorithms and modelling discussions are based on the introduction of a so-called reaction coordinate ξ, namely a smooth low-dimensional function which typically indices transition from one metastable state to another. For simplicity, let us assume that $\xi : \mathbb{R}^n \to \mathbb{T}$ has values in the one-dimensional torus (think of an angle in a molecule, which characterizes its conformation). Then, one may introduce probability measures associated to μ and ξ:

- The image $\xi * \mu$ of the measure μ by ξ, which is a probability measure on the torus $\mathbb{T}$, and which is also written as $\xi * \mu(dz) = \exp(-\beta F(z))\, dz$, $F : \mathbb{T} \to \mathbb{R}$ being then the so-called free energy associated to μ and ξ. Using the co-area formula, a formula for F is given by

$$F(z) = -\beta^{-1} \ln \int_{\Sigma(z)} Z^{-1} \exp(-\beta V(x)) \delta_{\xi(x)-z}(dx) \tag{8}$$

  where $\Sigma(z) = \{x \in \mathbb{R}^n, \xi(x) = z\}$ and $\delta_{\xi(x)-z}(dx)$ is a measure supported by $\Sigma(z)$ such that $dx = \delta_{\xi(x)-z}(dx)\, dz$.
- The family of conditional probability measures $\mu(\cdot|\xi(x) = z)$ with support $\Sigma(z)$, which are indexed by $z \in \mathbb{T}$ and defined as

$$\mu(dx|\xi(x) = z) = \frac{\exp(-\beta V(x)) \delta_{\xi(x)-z}(dx)}{\exp(-\beta F(z))}.$$

These two measures are completely defined through the conditioning formula: for any test functions: $\varphi : \mathbb{R}^n \to \mathbb{R}$ and $\psi : \mathbb{T} \to \mathbb{R}$,

$$\int_{\mathbb{R}^n} \varphi(\xi(x)) \psi(x) \mu(dx) = \int_{\mathbb{T}} \varphi(z) \int_{\Sigma(z)} \psi(x) \mu(dx|\xi(x) = z) \exp(-\beta F(z))\, dz.$$

For proofs, we refer for example to [36].

Let us assume that the measures $\mu(\cdot|\xi(x) = z)$ satisfy a LSI(ρ) (with ρ uniform in z), and that the measure $\xi * \mu = Z^{-1} \exp(-\beta F(z))\, dz$ satifies a LSI(r). In the spirit of the previous subsection, we will say that (this will be assumption (H1) below):

$$\text{“the metastability of the process } X_t \text{ is along } \xi\text{” if and only if } \rho \gg r. \tag{9}$$

A typical example of such a situation is given on Fig. 1, for which $\xi(x, y) = x$. In such a two-dimensional setting, notice that $F(x) = -\beta^{-1} \ln \int_{\mathbb{R}} Z^{-1} \exp(-\beta V(x, y))\, dy$ and $\mu(dy|\xi(x, y) = x) = \dfrac{\exp(-\beta V(x, y))\, dy}{\exp(-\beta F(x))}$.

It is possible to relate the LSI constant of the measure μ (namely R) to the LSI constants of the measures $\mu(\cdot|\xi(x) = z)$ (namely ρ) and $\xi * \mu$ (namely r), see [32].

Roughly speaking,

$$\text{If } r \text{ is small and } \rho \text{ is small, then } R \text{ is small.} \tag{10}$$

Moreover, if R is small and ξ is well chosen, then r is small but ρ may be very large. In such a case, the metastability of the process X_t is essentially encoded in the low-dimensional observable $\xi(X_t)$. It is then possible to use numerical and analytical techniques to accelerate the long-time convergence and circumvent the difficulties associated to metastability, as explained in the two next subsections. This could also be used to yield a definition of what a good reaction coordinate is: it is a (low-dimensional) function ξ such that ρ/r is as large as possible. Designing a numerical method which would look for the best ξ in this respect would be very interesting for practical applications.

Before we proceed, we provide a formula that will be useful below. By using the co-area formula, starting from (8), it is possible to check that the derivative of F writes:

$$F'(z) = \int f(x)\mu(dx|\xi(x) = z) \text{ with } f = \frac{\nabla V \cdot \nabla \xi}{|\nabla \xi|^2} - \beta^{-1}\text{div}\left(\frac{\nabla \xi}{|\nabla \xi|^2}\right). \tag{11}$$

In the simple case $\xi(x, y) = x$ mentioned above, the function f is simply $f = \partial_x V$.

2.3 A First Example: The Adaptive Biasing Force Technique

As explained above, one difficulty related to metastability is that the convergence (4) is very slow. In particular, this implies that it will be difficult to sample the canonical measure μ from a trajectory X_t. The fact that it is difficult to sample a multimodal measure is a well known problem shared with other fields than molecular dynamics. In statistics for example, Markov Chain Monte Carlo methods are very popular and similar sampling problems occur for Bayesian inference [10].

Under assumption (9), a natural importance sampling idea is the following. If the metastability of X_t is along ξ, it is sensible to try to remove these metastable features by changing the potential V to $V - F \circ \xi$, where F is the free energy (8) (and $F \circ \xi$ denotes the composition of F with ξ). Indeed, if we denote

$$\mu_F(dx) = Z_F^{-1} \exp(-\beta(V - F \circ \xi)(x))\, dx \tag{12}$$

the associated tilted measure, we clearly have $\xi * \mu_F = 1_{\mathbb{T}}(x)\, dx$ and $\mu_F(\cdot|\xi(x) = z) = \mu(\cdot|\xi(x) = z)$: the conditional measures remain the same, but the marginal along ξ is now a uniform measure, namely a very gentle measure, without any multimodality, and thus easy to sample. In other words, following (10), we may hope that the LSI constant of the tilted measure μ_F is much smaller than the one

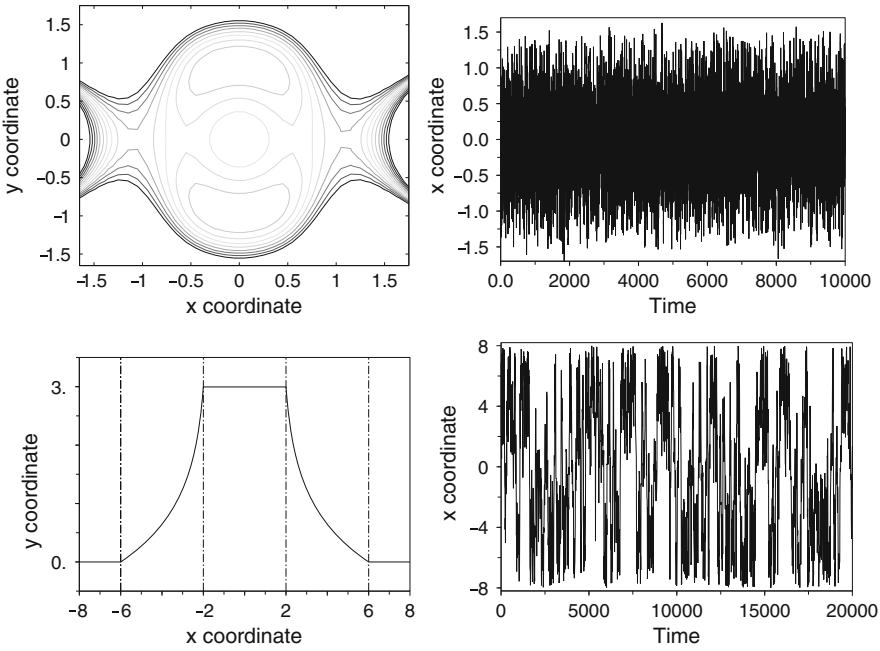


Fig. 2 See Fig. 1 for comparison. The reaction coordinate is $\xi(x, y) = x$. Above: In the energetic barrier case, (*left*) the 2 dimensional potential minus the free energy and (*right*) the x-component of the associated stochastic process. Below: In the entropic barrier case, (*left*) the free energy and (*right*) the x-component of the stochastic process simulated again with the free energy biased potential

of the original measure μ. As a numerical illustration of this fact, we present on Fig. 2 the equivalent of the trajectories presented on Fig. 1, using the biased potential $V - F \circ \xi$. We clearly observe much more transitions from left to right with the biased potential, both in the case of an energetic barrier and an entropic barrier.

The difficulty is of course that computing the free energy F is a challenge in itself (it corresponds to a sampling problem of, a priori, a similar complexity as the sampling of μ). The free energy is actually a quantity of great interest for practitioners in itself [9], and many methods have been designed to compute it [36]. Thus, biasing the measure using F does not seem to be such a good idea. The principle of adaptive biasing technique is to actually use, at a given time t, *an approximation* F_t of the free energy F (in view of the configurations visited so far) in order to bias the dynamics. In other words, instead of using the biased dynamics (which assumes that F is already known)

$$\begin{cases} dX_t = -\nabla(V - F \circ \xi)(X_t)\, dt + \sqrt{2\beta^{-1}}\, dW_t, \\ F'(z) = \mathbb{E}_\mu(f(X)|\xi(X) = z), \end{cases} \tag{13}$$

where the formula for $F'(z)$ is exactly (11) (here and in the following, $\mathbb{E}_\mu$ denotes an expectation taken with respect to the measure μ), one rather considers

$$\begin{cases} dX_t = -\nabla(V - F_t \circ \xi)(X_t)\, dt + \sqrt{2\beta^{-1}}\, dW_t, \\ F_t'(z) = \mathbb{E}(f(X_t)|\xi(X_t) = z). \end{cases} \tag{14}$$

The bottom line is that if X_t solution to (14) was at equilibrium instantaneously with respect to μ_{F_t} (using the notation (12)), then F_t' would be exactly F'. Of course, this instantaneous equilibrium assumption is wrong, but the hope is that F_t' learns on the fly (and eventually converges to) F'.

The dynamics (14) is the so called adaptive biasing force (ABF) process, and is one of the most efficient methods to compute free energy differences, see [13, 25] for the original idea. One way to understand such a method is that already visited states are penalized (the potential is flattened in these regions, or equivalently, the probability to visit them is increased) in order to force the stochastic process to visit new regions. There are many other techniques along these lines, see [34]. The hope is that, by forcing the system to visit all the possible values of ξ, the metastability of the original dynamics is completely overcome. This is somewhat in the spirit of simulated annealing or parallel tempering, where the temperature (which in some sense plays the role of ξ) of the system is changed in order to visit new regions.

What can be shown is that under the assumptions:

- (H1) the metastability of the process X_t is along ξ (see (9)),
- (H2) the cross derivative $\nabla_{\Sigma(z)} f$ is bounded (where $\nabla_{\Sigma(z)}$ denotes the gradient projected onto the tangent space to $\Sigma(z)$),

then the convergence of the ABF process (13) to equilibrium is much faster (basically exponential with rate $\beta^{-1}\rho$) compared to the convergence of the original process (1) to equilibrium (which is exponential with rate $\beta^{-1}R$, see (6)). In particular, F_t converges to the free energy F very fast if ξ is such that ρ is large (this is assumption (H1)). We refer to [35] for a precise mathematical statement. The proof is based on entropy techniques [2], and the idea of two-scale analysis for logarithmic Sobolev inequalities [22, 32].

We also refer to [33] for some refinements in the cases when ρ is only large for some values of z in the family of conditional probability measures $\mu(\cdot|\xi(x) = z)$ indexed by z (the so-called bi-channel situation). For practical aspects (discretization techniques and numerical illustrations), we refer to [8, 27, 34] and to [10] for an application of such techniques in the context of Bayesian inference in statistics.

Remark 3. In the long-time limit, F is obtained, and the measure sampled by (14) is not μ but μ_F. There are basically two ways to recover averages with respect to μ. First, as in standard importance sampling approaches, reweighting can be used:

$$\mathbb{E}_\mu(\varphi(X)) = \frac{\mathbb{E}_{\mu_F}(\varphi(X)\exp(-\beta F \circ \xi(X))}{\mathbb{E}_{\mu_F}(\exp(-\beta F \circ \xi(X))}.$$

For this idea to be efficient, the weights should not be too widespread (otherwise the variance may be large), which means that $\sup F - \inf F$ should not be too large, see [10] for a discussion of these aspects. Another idea is to use a conditioning approach:

$$\mathbb{E}_\mu(\varphi(X)) = \frac{\displaystyle\int_{\mathbb{T}} \mathbb{E}_\mu(\varphi(X)|\xi(X) = z) \exp(-\beta F(z))\, dz}{\displaystyle\int_{\mathbb{T}} \exp(-\beta F(z))\, dz}.$$

The conditional probabilities $\mathbb{E}_\mu(\varphi(X)|\xi(X) = z)$ can then be computed using either the fact that $\mathbb{E}_\mu(\varphi(X)|\xi(X) = z) = \mathbb{E}_{\mu_F}(\varphi(X)|\xi(X) = z)$ (so that the ABF process (14), in the longtime limit, can be used to compute them) or using dedicated techniques to sample the conditional probability measure $\mu(\cdot|\xi(x) = z)$, which should be easy under assumption (H1). Indeed, if ρ is large, in virtue of Proposition 1, the overdamped Langevin dynamics associated to the measure $\mu(\cdot|\xi(x) = z)$ should converge very fast to equilibrium. Such a dynamics is roughly speaking a projection of the original gradient dynamics (1) onto the submanifold $\Sigma(z)$. We refer to [11, 36, 37] for more information on such constrained sampling techniques.

2.4 A Second Example: Obtaining an Effective Dynamics on $\xi(X_t)$

If we are in the situation (9) where the metastability of X_t is along ξ, another idea is to try to derive an effective dynamics for $\xi(X_t)$, which would then be easy to simulate since it is low-dimensional, and hopefully associated with a smaller characteristic timescale than the original dynamics (1). The idea is that if the metastability is along ξ, then $\xi(X_t)$ should move much more slowly than the components of X_t which are along "directions orthogonal to ξ", so that some averaging should be possible along those directions. This is very much in the spirit of projection operator or Mori-Zwanzig techniques [20]. Below, we first derive an effective Markovian dynamics for $\xi(X_t)$ and then assess the quality of this effective dynamics by deriving quantitative bounds. These quantitative bounds are again obtained using logarithmic Sobolev inequalities and entropy computations.

The idea to derive an effective dynamics on $\xi(X_t)$ starts from a simple Itô calculus. If $(X_t)_{t\geq 0}$ satisfies (1), then

$$d\xi(X_t) = (-\nabla V \cdot \nabla\xi + \beta^{-1}\Delta\xi)(X_t)\, dt + \sqrt{2\beta^{-1}}|\nabla\xi(X_t)|\frac{\nabla\xi(X_t)}{|\nabla\xi(X_t)|} \cdot dW_t.$$

Note that $\int_0^t \frac{\nabla\xi(X_s)}{|\nabla\xi(X_s)|} \cdot dW_s$ is a one-dimensional Brownian motion. Of course, this is not a closed equation for the evolution of $\xi(X_t)$. It can be checked that if we consider

$$d\tilde{z}_t = \tilde{b}(t,\tilde{z}_t)\,dt + \sqrt{2\beta^{-1}}\tilde{\sigma}(t,\tilde{z}_t)\,dB_t$$

with

$$\tilde{b}(t,\tilde{z}) = \mathbb{E}\left((-\nabla V\cdot\nabla\xi + \beta^{-1}\Delta\xi)(X_t)\Big|\xi(X_t) = \tilde{z}\right)$$

and

$$\tilde{\sigma}^2(t,\tilde{z}) = \mathbb{E}\left(|\nabla\xi|^2(X_t)\Big|\xi(X_t) = \tilde{z}\right),$$

then, for all time $t \geq 0$, the law of the random variable $\xi(X_t)$ is *equal* to the law of the random variable $\tilde{z}_t$. Here, B_t denotes a one-dimensional Brownian motion. The difficulty of course is that $\tilde{b}$ and $\tilde{\sigma}$ are intractable numerically, since they are functions depending on time. A natural idea (following the intuition given at the beginning of this section) is that one could replace the conditional expectations defining $\tilde{b}$ and $\tilde{\sigma}^2$ by conditional expectations *at equilibrium* (i.e. with respect to the measure μ), namely:

$$dz_t = b(z_t)\,dt + \sqrt{2\beta^{-1}}\sigma(z_t)\,dB_t \tag{15}$$

with

$$b(z) = \mathbb{E}_\mu\left((-\nabla V\cdot\nabla\xi + \beta^{-1}\Delta\xi)(X)\Big|\xi(X) = z\right)$$

and

$$\sigma^2(z) = \mathbb{E}_\mu\left(|\nabla\xi|^2(X)\Big|\xi(X) = z\right).$$

For related approaches, we refer to [14, 40, 42].

Now that we have derived a Markovian dynamics (15) on z_t, which should be such that $(z_t)_{t\geq 0}$ is close to $(\xi(X_t))_{t\geq 0}$, a natural question is whether we can give some error estimate on some "distance between the two processes". What we have shown in [29] is that under the same assumptions (H1) and (H2) needed for the analysis of the longtime convergence of the ABF dynamics, the relative entropy of the law at time t of $\xi(X_t)$ with respect to the law at time t of z_t is bounded from above *uniformly in time* by a constant divided by ρ^2. Thus, the larger ρ (this is exactly (H1)), the closer these two probability measures are, for all times. We refer to [29, 30] for a precise mathematical statement. The proof uses basically the same ingredients as for the analysis of the long-time convergence of the ABF process.

Remark 4. The result we mention above only concerns the closeness of the *marginals in time* of the two processes, namely the laws, at times $t \geq 0$, of $\xi(X_t)$ and z_t. Of course, the stochastic processes $(\xi(X_t))_{t\geq 0}$ and $(z_t)_{t\geq 0}$ contain much more information than their marginal in times (think of time correlations, first time of escape from a well, typical paths to go from one well to another, etc...), which are of interest in the context of molecular dynamics simulation. A natural question is thus whether one can prove similar results on the *law of the paths*. This can indeed be done, under very close assumptions to (H1) and (H2), on finite time intervals. We refer to [31].

3 Quasi-stationary Distribution

In this second part, we would like to introduce another tool to study metastability, and to show in particular how this tool may be useful to analyze the parallel replica dynamics [48] which is a numerical method to efficiently generate *a trajectory* of a metastable dynamics.

Let us again consider the dynamics (1), and let us assume that we are given some partition of the state space, for example through an application

$$\mathscr{S} : \mathbb{R}^n \to \mathbb{N}$$

which to a given configuration x associates the number $\mathscr{S}(x)$ of the state in which x is. In some sense, $\mathscr{S}$ could be seen as an equivalent to the reaction coordinate ξ of the previous sections, but with discrete values. In the following, one should think of the states (for $k \in \mathbb{N}$)

$$W_k = \{x \in \mathbb{R}^n,\, \mathscr{S}(x) = k\}$$

as the metastable regions we mentioned in the introduction (think for example of the basins of attraction of the local minima of V). Typically, when X_t enters a state W_k, we would like it to stay in the state for a long time before it leaves the state. To formalize this idea, and to quantify the error introduced when this assumption is done in some algorithms, we rely on the notion of quasi-stationary distribution. In the following, we assume that the states W_k are smooth bounded connected subsets of $\mathbb{R}^n$.

3.1 Definition and Two Basic Properties of the Quasi-stationary Distribution

In this section, we consider a given state W_k for a fixed index k, and we denote by W the interior of W_k. Let us introduce the notion of the quasi-stationary distribution associated to the well W. We refer to [6, 12, 16–18, 39, 41, 43, 46] for general introductions to the quasi-stationary distribution.

Definition 2. The quasi-stationary distribution (QSD) associated to the dynamics (1) and the state W is defined as a measure ν with support in W and such that $\forall t > 0, \forall A \subset W$,

$$\nu(A) = \frac{\displaystyle\int_W \mathbb{P}(X_t^x \in A,\, t < T_W^x)\, \nu(dx)}{\displaystyle\int_W \mathbb{P}(t < T_W^x)\, \nu(dx)},$$

where X_t^x denotes the solution to (1) such that $X_0 = x$ and $T_W^x = \inf\{t \geq 0,\, X_t^x \notin W\}$.

In other words, if X_0 is distributed according to ν and if $(X_s)_{s\geq 0}$ solution to (1) has not left the state W on the interval $[0, t]$, then X_t is also distributed according to ν.

The QSD enjoys two other properties which are very important in practice, given in the two following propositions. To present these two properties, we need to state an intermediate result. Let $L = -\nabla V \cdot \nabla + \beta^{-1}\Delta$ be the infinitesimal generator associated to (1). Then the density u_1 of ν with respect to μ (namely such that $\nu(dx) = u_1(x)\mu(dx)$) is the first eigenfunction of L with zero Dirichlet boundary condition on ∂W (and with normalization $\int_W u_1 d\mu = 1$). In other words:

$$u_1 \in \operatorname{argmin}_{u\in H^1_{\mu,0}(W)} \frac{\displaystyle\int_W |\nabla u|^2 \, d\mu}{\displaystyle\int_W u^2 \, d\mu} \tag{16}$$

and

$$\begin{cases} Lu_1 = -\lambda_1 u_1 \text{ in } W, \\ u_1 = 0 \text{ on } \partial W. \end{cases} \tag{17}$$

Here $H^1_{\mu,0}(W) = \{u : W \to \mathbb{R}, \int_W \left(|\nabla u|^2 + u^2\right) d\mu < \infty \text{ and } u = 0 \text{ on } \partial W\}$. The operator L can be shown to be negative, self-adjoint on L^2_μ, and with a discrete spectrum $(-\lambda_1, -\lambda_2, \ldots, -\lambda_n, \ldots)$ (the eigenvalues are in decreasing order and counted with multiplicity) and associated eigenfunctions $(u_1, u_2, \ldots, u_n, \ldots)$ in $H^1_{\mu,0}(W)$. The first eigenstate can be shown to be non-degenerate ($\lambda_2 > \lambda_1 > 0$), and the first eigenfunction does not vanish on W, and can thus be assumed to be positive ($u_1 > 0$ on W). Using this spectral decomposition of the infinitesimal generator, and this characterization of the QSD, one can show the following results.

Proposition 2. *Let X_0 be distributed according to a distribution with support in W, and let us consider $(X_t)_{t\geq 0}$ solution to (1). Then, the law of X_t conditionally to the fact that the process remained in W up to time t converges to the QSD ν: for any test function $\varphi : W \to \mathbb{R}$,*

$$\lim_{t\to\infty} \mathbb{E}(\varphi(X_t) \,|\, t < T_W) = \int_W \varphi d\nu.$$

Moreover, the rate of convergence is exponential:

$$\sup_{\varphi\in L^\infty(W), \|\varphi\|_{L^\infty(W)}\leq 1} \left| \mathbb{E}(\varphi(X_t) \,|\, t < T_W) - \int_W \varphi d\nu \right| \leq C \exp(-(\lambda_2 - \lambda_1)t), \tag{18}$$

where, as explained above, $-\lambda_1 > -\lambda_2$ are the two first eigenvalues of the operator $L = -\nabla V \cdot \nabla + \beta^{-1}\Delta$ considered on W with zero Dirichlet boundary conditions on ∂W.

The proof of this proposition is based on a spectral decomposition of the solution to the associated Feynman-Kac partial differential equation, and the simplest proof actually requires the law of X_0 to have a density with respect to μ which is in L^2_μ.

This Proposition has two consequences. First, if the process X_t enters a state and stays sufficiently long in the state, then its marginal in time is close to the QSD. Second, one could also think of considering the following interacting particle system, which samples the law of X_t conditionally to the fact that $t < T_W$:

- Consider N initial conditions $(X_0^n)_{1\leq n\leq N}$ distributed independently according to a given law with support in W;
- Let them evolve according to (1), driven by independent Brownian motions;
- Each time a replica leaves the state, it is killed, another replica is duplicated and the new walker then evolves again independently of the others.

This is the so-called Flemming-Viot process [5, 17, 21, 38]. In the limit of infinitely many replicas, the replicas are distributed according to the law of X_t conditionally to the fact that $t < T_W$, so that, in the long-time limit, the replicas are distributed according to the QSD (according to Proposition 2).

Finally, another crucial property of the QSD is the following:

Proposition 3. *Let X_0 be distributed according to the QSD, and let us consider $(X_t)_{t\geq 0}$ solution to (1). Let us recall that $T_W = \inf\{t \geq 0, X_t \notin W\}$. Then,*

- *The law of the exit time T_W is exponential with parameter λ_1;*
- *The law of the exit point X_{T_W} is $\left(-\frac{1}{\beta\lambda_1}\frac{\partial u_1}{\partial n}\exp(-\beta V)\right) d\lambda_{\partial W}$ where n denotes the unit outward normal to W and $\lambda_{\partial W}$ is the Lebesgue measure on ∂W;*
- *The two random variables T_W and X_{T_W} are independent.*

For proofs of these well-known properties of the QSD in our precise setting, we refer to [28].

In this context, one could state that the region W is a metastable state for the dynamics (1) if the typical time it takes to leave W is large compared to the typical time it takes to reach the QSD (which is $1/(\lambda_2 - \lambda_1)$ according to (18)), namely:

$$\text{The probability } \mathbb{P}\left(T_W < \frac{1}{\lambda_2 - \lambda_1}\right) \text{ is close to zero.} \tag{19}$$

In some sense, this characterization (19) of metastability is the counterpart, in terms of QSD, of the two characterizations (7) and (9) we introduced above in terms of LSI. The difficulty to completely formalize (19) is that the law of the initial condition $X_0 \in W$ should be defined to make precise the law of T_W. For example, if X_0 is distributed according to the QSD, T_W is exponential with parameter λ_1 and thus $\mathbb{P}\left(T_W < \frac{1}{\lambda_2-\lambda_1}\right) = 1 - \exp(-\lambda_1/(\lambda_2 - \lambda_1))$ so that W is metastable if $\lambda_1 \ll \lambda_2 - \lambda_1$. As in the LSI case where (9) could be used to define a good reaction coordinate, the characterization (19) of metastability in terms of the QSD could be useful to define what a good partition of the configurational space (namely a good function $\mathscr{S}$) is.

3.2 A First Application: Analysis of the Parallel Replica Dynamics

The notion of QSD can be used to analyze an algorithm called the parallel replica dynamics which has been introduced by A.F. Voter in [48], and which is based on some Markovianity assumption, as will become clear below. The QSD is a way to quantify the error introduced by this Markovianity assumption, and to explain in which context the algorithm is efficient. The aim of the parallel replica dynamics is to generate very efficiently a process $(S_t)_{t\geq 0}$ with values in $\mathbb{N}$, and which is close to $(\mathscr{S}(X_t))_{t\geq 0}$. Indeed, in many cases, one is not interested in the details of the dynamics of X_t: only the hopping events from one state to another are of interest. One important requirement is that the two stochastic processes $(S_t)_{t\geq 0}$ and $(\mathscr{S}(X_t))_{t\geq 0}$ should be close in terms of *the law of the trajectories* (not only the time marginals, for example, see also Remark 4 above).

Let us first describe the algorithm. This is a three stage algorithm. In the *decorrelation step* a reference walker evolves according to (1), up to a time it stayed sufficiently long in the same state. More precisely, a time denoted τ_{corr} is introduced, and one proceeds to the next stage at a time t_0 if and only if $\mathscr{S}(X_t) = k_0$ is constant over the time interval $[t_0 - \tau_{\text{corr}}, t_0]$. During all this step, S_t is by definition $\mathscr{S}(X_t)$ so that no error is introduced. Let us assume that the decorrelation step has been successful, and let us proceed to the *dephasing step*. It consists in introducing N replicas of the reference walker in the state W_{k_0} and to let them evolve (in parallel on different CPU cores) sufficiently long, conditionally to the fact that they do not leave the state, and to retain their final position. In other words, the dephasing step is basically a realization of the Fleming-Viot process introduced in the previous section. During this stage which is of course done in parallel, the process S_t is not evolved. Finally, the speed up comes from the last step called the *parallel step*. It consists in letting all the walkers evolve independently and in parallel on different CPU cores from the initial conditions obtained in the previous dephasing step. Then the first escape event is detected, namely

$$n_0 = \text{argmin}_{n\in\{1,\dots,N\}}\{T_W^n\}$$

where T_W^n is the first escape time from W for the n-th replica. Then, $S_t = k_0$ over the time interval $[t_0, t_0 + NT_W^{n_0}]$, and one proceeds to a new decorrelation step, the reference walker starting from the exit point $X_{T_W^{n_0}}^{n_0}$. The speed-up of course comes from the fact that we consider only the first escape event among the N walkers. As will become clear below, this event occurs in a time N times smaller than the time it would take for a single walker to leave the state.

Let us now discuss the error analysis of this procedure. A first remark is that at the end of the decorrelation step, if τ_{corr} has been chosen sufficiently large (namely $\tau_{\text{corr}} \gg \dfrac{1}{\lambda_2 - \lambda_1}$, according to (18)), it is reasonable to assume that X_{t_0} is approximately distributed according to the QSD. Thus, according to Proposition 3,

the time it still remains to go out of the state W_{k_0} is exponentially distributed, and independent of the exit point. Then, the aim of the dephasing step is clear: one wants to obtain N initial conditions independently and identically distributed according to the QSD. The parallel step is thus fully justified. Indeed, concerning the time spent in the state k_0, since $(T_W^1, \ldots, T_W^N)$ are N i.i.d. exponential random variables, $T_W^{n_0} = \min_n T_W^n$ is also an exponential random variable and $N T_W^{n_0}$ has the same law as T_W^1: this explains why the simulation clock is advanced by the amount of time $N T_W^{n_0}$ at the end of the parallel step. Moreover, concerning the exit point, since the exit time and the exit point are independent random variables when starting from the QSD, considering $X^{n_0}_{T_W^{n_0}}$ as the exit point is correct in terms of distribution (it has the same law as $X^1_{T_W^1}$).

In summary, the crucial parameter is $\tau_{\rm corr}$, which is used in the decorrelation step. The error which is made by one iteration of the algorithm can be formalized by considering:

$$e(t) = \sup_{f:\mathbb{R}_+\times\partial W\to\mathbb{R}, \|f\|_{L^\infty}\le 1} \left|\mathbb{E}(f(T_W - t, X_{T_W})|T_W \ge t) - \mathbb{E}_\nu(f(T_W, X_{T_W}))\right|,$$

where $\mathbb{E}_\nu$ here denotes an expectation over functionals of X_t, the initial condition X_0 being distributed according to ν. In words, $e(\tau_{\rm corr})$ measures the difference of what would have been the law of the couple of random variables (exit time, exit point) if the simulation of the reference walker would have been continued after the time $\tau_{\rm corr}$, compared to the law of the same couple of random variables, if we assume that the reference walker is distributed according to the QSD. A slight adaptation of Proposition 2 above shows that $e(t) \le C\exp(-(\lambda_2 - \lambda_1)t)$ so that $\tau_{\rm corr}$ should be chosen larger than $1/(\lambda_2 - \lambda_1)$. The constant C here depends on the law of the initial condition in the state. On the other hand, if $\tau_{\rm corr}$ is much larger than the typical time it takes to leave the state, the decorrelation step has no chance to be successful, and no speed-up is achieved. With these two requirements on $\tau_{\rm corr}$, it thus becomes clear that this algorithm is efficient if most of the states are metastable, in the sense of (19).

In conclusion, the interest of the QSD in this context is twofold: (i) it enables to understand how large $\tau_{\rm corr}$ should be in order not to introduce too much error in one iteration of the algorithm and (ii) it helps to define the assumptions required for the algorithm to be efficient.

3.3 A Second Application: Going from Continuous State Space Dynamics to Kinetic Monte Carlo Models

The notion of QSD could also be useful in order to formalize the construction of discrete state space Markov models (so called kinetic Monte Carlo models [49] in the context of molecular dynamics) from continuous state space Markov models such as (1). Let us recall that a stochastic process S_t with values in $\mathbb{N}$ is Markovian

if and only if: (M1) the list of visited states (forgetting about time) is Markovian and (M2) once S_t takes a new value (it enters a new state), the time it takes to leave this state is exponentially distributed and independent of the next visited state.

There are basically two approaches to build such a connection. In the so-called milestoning approach [15,45], one considers some disjoint subsets (the milestones) of the configuration space (think of small balls around the local minima of the potential V) and one considers the last milestone visited by X_t. The interest of this approach is that in the limit of very small subsets, the first requirement above (M1) is naturally satisfied. On the other hand, satisfying (M2) is more involved, and requires some assumptions related to the metastability of the process. Here metastability basically means that the time spent outside the milestones is very small compared to the time spent in the milestones, see [3,4]. The main drawback of this approach is that if the milestones are too small, the process X_t spends a significant time outside of the milestones, so that the stochastic process built as "the last visited milestone" may not be a sufficiently fine coarse-grained description of the original process X_t, in order to extract useful macroscopic information (change of conformation of a molecular system, for example). Of course, this depends a lot on the system at hand.

The second natural approach which has been followed in the previous section and by many authors [26,44,49] is to consider a full partition of the state space, and to consider at a given time t, in which state the process is. With the previous notation, it thus consists in considering $\mathscr{S}(X_t)$. Again, the process $\mathscr{S}(X_t)$ has no reason to be Markovian. The approximation suggested by the approach outlined above is to introduce a Markov process S_t as follows: when S_t jumps to a new value k_0 (one can imagine that the underlying process X_t just enters a new state W_{k_0}), the time it takes to leave the value k_0 is exponentially distributed with parameter λ_1, and, independently, the next visited state is drawn according to the exit point distribution $\left(-\frac{1}{\beta\lambda_1}\frac{\partial u_1}{\partial n}\exp(-\beta V)\right) d\lambda_{\partial W}$. Here, we used the notation of the previous section: in particular, (λ_1, u_1) are the first eigenvalue and eigenfunction of the operator L on W, with zero Dirichlet boundary condition on ∂W. This procedure would be exact if, as soon as X_t entered a new state, it would immediately be distributed according to the QSD. The error introduced by this coarse-grained description is thus related to the metastability of the original process, namely to the fact that when it enters a new state, it reaches the QSD before leaving the state (this is (19)). Contrary to the previous results presented in this paper, we have not yet fully formalized these ideas from a mathematical viewpoint. Notice that in the approach we propose here, the kernel of the approximating Markov process is computed using the QSD as an initial distribution in a given state, in contrast to what can be found usually in the literature, namely starting from the canonical measure μ restricted to the state. The interest of starting with the QSD is that the underlying assumptions ruling a Markov process (see (M1) and (M2) above) are automatically satisfied.

Acknowledgements I would like to thank my co-authors on these subjects (Chris Chipot, Nicolas Chopin, Giovanni Ciccotti, Brad Dickson, Benjamin Jourdain, Claude Le Bris, Frédéric Legoll, Mitch Luskin, Kimiya Minoukadeh, Stefano Olla, Danny Perrez, Mathias Rousset, Raphael Roux,

Gabriel Stoltz and Eric Vanden-Einjden) as well as Félix Otto who introduced me to the so-called two-scale analysis for logarithmic Sobolev inequalities and Art Voter for very useful and inspiring discussions. This work is supported by the Agence Nationale de la Recherche, under grant ANR-09-BLAN-0216-01 (MEGAS).

References

1. C. Ané, S. Blachère, D. Chafaï, P. Fougères, I. Gentil, F. Malrieu, C. Roberto, and G. Scheffer *Sur les inégalités de Sobolev logarithmiques*. Société Mathématique de France, 2000. In French.
2. A. Arnold, P. Markowich, G. Toscani, and A. Unterreiter. On convex Sobolev inequalities and the rate of convergence to equilibrium for Fokker-Planck type equations. *Comm. Part. Diff. Eq.*, 26:43–100, 2001.
3. A. Bovier, M. Eckhoff, V. Gayrard, and M. Klein. Metastability in reversible diffusion processes. I. sharp asymptotics for capacities and exit times. *J. Eur. Math. Soc. (JEMS)*, 6:399–424, 2004.
4. A. Bovier, V. Gayrard, and M. Klein. Metastability in reversible diffusion processes. II. precise asymptotics for small eigenvalues. *J. Eur. Math. Soc. (JEMS)*, 7:69–99, 2004.
5. K. Burdzy, R. Holyst, and P. March. A Fleming-Viot particle representation of the Dirichlet Laplacian. *Communications in Mathematical Physics*, 214(3):679–703, 2000.
6. P. Cattiaux, P. Collet, A. Lambert, S. Martínez, S. Méléard, and J. San Martín. Quasi-stationary distributions and diffusion models in population dynamics. *Ann. Probab.*, 37(5):1926–1969, 2009.
7. P. Cattiaux and A. Guillin. Functional inequalities via Lyapunov conditions, 2010. To appear in SMF, Collections Congrès et Seminaires. Available at http://arxiv.org/abs/1001.1822.
8. C. Chipot and T. Lelièvre. Enhanced sampling of multidimensional free-energy landscapes using adaptive biasing forces. *SIAM J. Appl. Math.*, 71(5):1673–1695, 2011.
9. C. Chipot and A. Pohorille, editors. *Free Energy Calculations*, volume 86 of *Springer Series in Chemical Physics*. Springer, 2007.
10. N. Chopin, T. Lelièvre, and G. Stoltz. Free energy methods for Bayesian inference: efficient exploration of univariate Gaussian mixture posteriors. *Stat. Comput.*, 22(4):897–916, 2012.
11. G. Ciccotti, T. Lelièvre, and E. Vanden-Eijnden. Projection of diffusions on submanifolds: Application to mean force computation. *Commun. Pur. Appl. Math.*, 61(3):371–408, 2008.
12. P. Collet, S. Martínez, and J. San Martín. Asymptotic laws for one-dimensional diffusions conditioned to nonabsorption. *Ann. Probab.*, 23(3):1300–1314, 1995.
13. E. Darve and A. Porohille. Calculating free energy using average forces. *J. Chem. Phys.*, 115:9169–9183, 2001.
14. W. E and E. Vanden-Eijnden. Metastability, conformation dynamics, and transition pathways in complex systems. In *Multiscale modelling and simulation*, volume 39 of *Lect. Notes Comput. Sci. Eng.*, pages 35–68. Springer, Berlin, 2004.
15. A.K. Faradjian and R. Elber. Computing time scales from reaction coordinates by milestoning. *J. Chem. Phys.*, 120(23):10880–10889, 2004.
16. P.A. Ferrari, H. Kesten, S. Martinez, and P. Picco. Existence of quasi-stationary distributions. a renewal dynamical approach. *Ann. Probab.*, 23(2):511–521, 1995.
17. P.A. Ferrari and N. Maric. Quasi-stationary distributions and Fleming-Viot processes in countable spaces. *Electronic Journal of Probability*, 12, 2007.
18. P.A. Ferrari, S. Martinez, and J. San Martin. Phase transition for absorbed Brownian motion. *J. Stat. Physics.*, 86(1/2):213–231, 1996.
19. M.I. Freidlin and A.D. Wentzell. *Random Perturbations of Dynamical Systems*. Springer-Verlag, 1984.

20. D. Givon, R. Kupferman, and A. Stuart. Extracting macroscopic dynamics: model problems and algorithms. *Nonlinearity*, 17(6):R55–R127, 2004.
21. I. Grigorescu and M. Kang. Hydrodynamic limit for a Fleming-Viot type system. *Stoch. Proc. Appl.*, 110(1):111–143, 2004.
22. N. Grunewald, F. Otto, C. Villani, and M.G. Westdickenberg. A two-scale approach to logarithmic Sobolev inequalities and the hydrodynamic limit. *Ann. Inst. H. Poincaé Probab. Statist.*, 45(2):302–351, 2009.
23. B. Helffer and F. Nier. *Hypoelliptic Estimates and Spectral Theory for Fokker-Planck Operators and Witten Laplacians*, volume 1862 of *Lecture Notes in Mathematics*. Springer, 2005.
24. B. Helffer and F. Nier. Quantitative analysis of metastability in reversible diffusion processes via a Witten complex approach: the case with boundary. *Mémoires de la SMF*, 105, 2006.
25. J. Hénin and C. Chipot. Overcoming free energy barriers using unconstrained molecular dynamics simulations. *J. Chem. Phys.*, 121:2904–2914, 2004.
26. W. Huisinga, S. Meyn, and C. Schütte. Phase transitions and metastability in markovian and molecular systems. *Ann. Appl. Prob.*, 14(1):419–158, 2004.
27. B. Jourdain, T. Lelièvre, and R. Roux. Existence, uniqueness and convergence of a particle approximation for the adaptive biasing force process. *ESAIM-Math. Model. Num.*, 44:831–865, 2010.
28. C. Le Bris, T. Lelièvre, M. Luskin, and D. Perez. A mathematical formalization of the parallel replica dynamics, Monte Carlo Methods Appl., 18(2):119–146, 2012.
29. F. Legoll and T. Lelièvre. Effective dynamics using conditional expectations. *Nonlinearity*, 23:2131–2163, 2010.
30. F. Legoll and T. Lelièvre. *Some remarks on free energy and coarse-graining*, volume 82 of *Lecture Notes in Computational Science and Engineering*, pages 279–329. Springer, 2012.
31. F. Legoll, T. Lelièvre, and S. Olla. Effective dynamics for the overdamped langevin equation: trajectorial error estimates, 2012. In preparation.
32. T. Lelièvre. A general two-scale criteria for logarithmic Sobolev inequalities. *J. Funct. Anal.*, 256(7):2211–2221, 2009.
33. T. Lelièvre and K. Minoukadeh. Long-time convergence of an adaptive biasing force method: the bi-channel case. *Archive for Rational Mechanics and Analysis*, 202(1):1–34, 2011.
34. T. Lelièvre, M. Rousset, and G. Stoltz. Computation of free energy profiles with adaptive parallel dynamics. *J. Chem. Phys.*, 126:134111, 2007.
35. T. Lelièvre, M. Rousset, and G. Stoltz. Long-time convergence of an adaptive biasing force method. *Nonlinearity*, 21:1155–1181, 2008.
36. T. Lelièvre, M. Rousset, and G. Stoltz. *Free energy computations: A mathematical perspective*. Imperial College Press, 2010.
37. T. Lelièvre, M. Rousset, and G. Stoltz. Langevin dynamics with constraints and computation of free energy differences. *Math. Comput.*, 81(280):2071–2125, 2012.
38. J.U. Löbus. A stationary Fleming-Viot type Brownian particle system. *Mathematische Zeitschrift*, 263(3):541–581, 2008.
39. P. Mandl. Spectral theory of semi-groups connected with diffusion processes and its application. *Czechoslovak Math. J.*, 11 (86):558–569, 1961.
40. L. Maragliano, A. Fischer, E. Vanden-Eijnden, and G. Ciccotti. String method in collective variables: minimum free energy paths and isocommittor surfaces. *J. Chem. Phys.*, 125:024106, 2006.
41. S. Martínez and J. San Martín. Classification of killed one-dimensional diffusions. *Ann. Probab.*, 32(1A):530–552, 2004.
42. G.A. Pavliotis and A.M. Stuart. *Multiscale methods: averaging and homogenization*. Springer, 2007.
43. R.G. Pinsky. On the convergence of diffusion processes conditioned to remain in a bounded region for large time to limiting positive recurrent diffusion processes. *Ann. Probab.*, 13(2):363–378, 1985.

44. M. Sarich, F. Noé, and C. Schütte. On the approximation quality of Markov state models. *Multiscale Model. Simul.*, 8(4):1154–1177, 2010.
45. C. Schütte, F. Noé, Jianfeng Lu, M. Sarich, and E. Vanden-Eijnden. Markov state models based on milestoning. *J. Chem. Phys.*, 134(20):204105, 2011.
46. D. Steinsaltz and S.N. Evans. Quasi-stationary distributions for one-dimensional diffusions with killing. *Trans. Amer. Math. Soc.*, 359(3):1285–1324, 2007.
47. C. Villani. Hypocoercivity. *Memoirs Amer. Math. Soc.*, 202, 2009.
48. A.F. Voter. Parallel replica method for dynamics of infrequent events. *Phys. Rev. B*, 57(22):R13 985, 1998.
49. A.F. Voter. *Radiation Effects in Solids*, chapter Introduction to the Kinetic Monte Carlo Method. Springer, NATO Publishing Unit, 2005.

On the Reliability of Error Indication Methods for Problems with Uncertain Data

I. Anjam, O. Mali, P. Neittaanmäki, and S. Repin

Abstract This paper is concerned with studying the effects of uncertain data in the context of error indicators, which are often used in mesh adaptive numerical methods. We consider the diffusion equation and assume that the coefficients of the diffusion matrix are known not exactly, but within some margins (intervals). Our goal is to study the relationship between the magnitude of uncertainty and reliability of different error indication methods. Our results show that even small values of uncertainty may seriously affect the performance of all error indicators.

1 Introduction

In problems related to partial differential equations, it is usually assumed that data of the problem are known exactly. However, quite often the data at hand is not complete. In many problems, the data is uncertain within some intervals. Material functions, geometrical data, and boundary conditions may include uncertainties, which arise due to incomplete knowledge on the model.

Studying the effects caused by uncertain data gained the attention of researchers later than analysis of fully determined problems. The probabilistic approach is based in studying stochastic partial differential equations (see, e.g., [12]). Another approach (the so-called "worst case scenario method") has been analyzed in [5].

I. Anjam (✉) · O. Mali · P. Neittaanmäki
University of Jyväskylä, Department of Mathematical Information Technology,
P.O. Box 35 (Agora), FI-40014, Jyväskylä, Finland
e-mail: immanuel.anjam@jyu.fi; olli.mali@jyu.fi; pekka.neittaanmaki@jyu.fi

S. Repin
V. A. Steklov Institute of Mathematics in St. Petersburg Fontanka 27, RU-191024,
St. Petersburg, Russia
e-mail: repin@pdmi.ras.ru

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_84,

In [6–8], two-sided estimates of the radius of the solution set were obtained for the reaction-diffusion problem. These estimates provide information on the *accuracy limit* generated by the uncertainty. These estimates are derived with the help of functional a posteriori estimates (for a consequent exposition of the theory see [9, 11]).

In this paper, we study the diffusion equation with uncertainty in the diffusion matrix. We assume that the uncertainty is of the form *mean value* $\pm$ *variation*, which is typical for engineering measurements.

Our goal is to study how incomplete knowledge about the coefficients of diffusion (physical parameters) impact the reliability of error indication. We have tested several commonly used indicators with the paradigm of a simple elliptic problem. The results show that the reliability of error indicators seriously depend on the rank of uncertainty encompassed in the diffusion matrix.

2 Problem Definition and Notation

Let $\Omega \in \mathbb{R}^d$ be a bounded and connected domain with Lipschitz continuous boundary $\partial\Omega$. By $H^1(\Omega)$ we denote the Sobolev space of scalar valued functions with square summable generalized derivatives. $H_0^1(\Omega)$ is the subspace of $H^1(\Omega)$ containing the functions vanishing on $\partial\Omega$. For vector valued functions, we also define the space $H(\mathrm{div}, \Omega) := \{w \in L_2(\Omega, \mathbb{R}^d) \mid \mathrm{div}\, w \in L_2(\Omega)\}$.

We consider the simplest elliptic problem: find $u \in H_0^1(\Omega)$ such that

$$-\mathrm{div}\, A\nabla u = f \qquad \text{in } \Omega, \tag{1}$$

$$u = 0 \qquad \text{on } \partial\Omega, \tag{2}$$

where $f \in L_2(\Omega)$. Assume that the coefficients are not fully known, i.e., the information that we really possess is that $A \in \mathscr{D}$, where

$$\mathscr{D} := \{A \in L_\infty(\Omega, \mathbb{M}^{d\times d}) \mid A = A_0 + \delta\Psi,\ \|\Psi\|_{L_\infty(\Omega,\mathbb{M}^{d\times d})} \le 1\},$$

where $\mathbb{M}^{d\times d}$ is the space of symmetric matrices, $A_0 \in L_\infty(\Omega, \mathbb{M}^{d\times d})$ is the known "mean" matrix, and $\delta \ge 0$ is the magnitude of variations. In other words, we assume that A belongs to a set generated by limited perturbations of the "mean" data.

The generalized statement of (1) and (2) consists of finding $u \in H_0^1(\Omega)$ such that

$$\int_\Omega A\nabla u \cdot \nabla w dx = \int_\Omega f w dx, \qquad \forall w \in H_0^1(\Omega).$$

We assume that

$$\underline{c}|\xi|^2 \ \le\ A_0\xi \cdot \xi \ \le\ \overline{c}|\xi|^2,$$

where $0 < \underline{c} \leq \overline{c}$. Thus, the "mean" problem is elliptic and has a unique solution u_0. The condition

$$0 \leq \delta < \underline{c} \tag{3}$$

guarantees that the perturbed problem remains elliptic, and possesses a unique solution u with any $A \in \mathscr{D}$. With this condition the "solution mapping" $\mathscr{S} : \mathscr{D} \to H_0^1$ is well defined. The solution set generated by the uncertain data will be referred to as the set $\mathscr{S}(\mathscr{D}) \subset H_0^1$.

The standard L_2-norm $\|v\|_{2;\Omega}$ is denoted by $\|v\|_\Omega$. We also introduce the weighed L_2-norm for vector valued functions w:

$$\|w\|_{2,\mu;\Omega} = \|w\|_{\mu;\Omega} := \int_\Omega \mu w \cdot w dx.$$

Using this notation, for each $A \in \mathscr{D}$ we have the energy norm $\|\nabla v\|_{A;\Omega}$.

By $\mathscr{T}_h$ we denote the partition of the domain Ω to the union of non-overlapping triangles. An element in $\mathscr{T}_h$ is denoted by T. For any $T \in \mathscr{T}_h$, $\mathscr{E}(T)$ denotes edges of T, and $\mathscr{N}(T)$ the nodes. The sets

$$\mathscr{E}_h := \bigcup_{T \in \mathscr{T}_h} \mathscr{E}(T) \quad \text{and} \quad \mathscr{N}_h := \bigcup_{T \in \mathscr{T}_h} \mathscr{N}(T)$$

contain all edges and all nodes of $\mathscr{T}_h$, respectively. For the sake of convenience, we also define the set of edges which approximate the boundary of the domain by $\mathscr{E}_{h,\partial\Omega} := \{E \in \mathscr{E}_h \mid E \,\widetilde{\subset}\, \partial\Omega\}$. The sets

$$\omega_E := \bigcup_{E \in \mathscr{E}(T')} T' \quad \text{and} \quad \omega_X := \bigcup_{X \in \mathscr{N}(T')} T'$$

define patches of elements associated with a given edge $E \in \mathscr{E}_h$ and node $X \in \mathscr{N}_h$, respectively.

For every $E \in \mathscr{E}_h$, we assign a unit vector n_E, which it is orthogonal to E.

Henceforth, the symbol $|\cdot|$ is used to denote area of a domain or length of an edge. The number of elements in a set is denoted by $\#(\cdot)$ and the diameter of $T \in \mathscr{T}_h$ is denoted by h_T.

3 Error Indication

In our analysis, we consider small disturbances of the matrix A of the form

$$A = A_0 + \delta B,$$

where the magnitude of variations δ satisfies the condition (3), and B is a symmetric 2×2-matrix. We note that since the amount of matrixes contained in $\mathscr{D}$ is much

Table 1 The values of Θ with two digit accuracy (example (10))

a $\mathbb{E}_X$ node averaging (4)

N_{elem}	N_{dof}	δ					
		0.005	0.01	0.02	0.03	0.04	0.05
800	441	0.15	0.28	0.52	0.68	0.80	0.85
3200	1681	0.30	0.53	0.80	0.89	0.88	0.98
12800	6561	0.53	0.80	0.88	1	1	1
51200	25921	0.80	0.88	1	1	1	1
115200	58081	0.89	1	1	1	1	1

b $\mathbb{E}_E$ edge averaging (5)

δ					
0.005	0.01	0.02	0.03	0.04	0.05
0.16	0.27	0.50	0.67	0.75	0.84
0.28	0.51	0.77	0.87	0.89	0.87
0.51	0.77	0.89	0.95	1	1
0.77	0.89	1	1	1	1
0.88	0.96	1	1	1	1

c $\mathbb{E}_{RF}$ residual, full (6)

N_{elem}	N_{dof}	δ					
		0.005	0.01	0.02	0.03	0.04	0.05
800	441	0.28	0.41	0.65	0.80	0.90	0.94
3200	1681	0.42	0.64	0.89	0.95	0.96	0.96
12800	6561	0.65	0.89	0.96	0.97	1	1
51200	25921	0.89	0.96	1	1	1	1
115200	58081	0.95	0.97	1	1	1	1

d $\mathbb{E}_{RJ}$ residual, jump (7)

δ					
0.005	0.01	0.02	0.03	0.04	0.05
0.16	0.26	0.50	0.68	0.73	0.88
0.28	0.51	0.77	0.90	0.91	0.89
0.50	0.77	0.90	0.89	0.99	1
0.76	0.90	0.99	1	1	1
0.89	0.88	1	1	1	1

e $\mathbb{E}_{GA}$ global averaging (8)

N_{elem}	N_{dof}	δ					
		0.005	0.01	0.02	0.03	0.04	0.05
800	441	0.15	0.25	0.51	0.66	0.77	0.84
3200	1681	0.29	0.51	0.78	0.89	0.91	0.88
12800	6561	0.51	0.77	0.90	0.96	1	1
51200	25921	0.77	0.89	1	1	1	1
115200	58081	0.88	0.96	1	1	1	1

f $\mathbb{E}_F$ functional maj. (9)

δ					
0.005	0.01	0.02	0.03	0.04	0.05
0.15	0.26	0.51	0.66	0.76	0.84
0.28	0.51	0.78	0.89	0.89	0.87
0.51	0.77	0.89	0.96	1	1
0.77	0.89	1	1	1	1
0.88	0.96	1	1	1	1

larger than those representable in such a form, the sensitivity of error indicators with respect to data uncertainty is even higher than indicated on Table 1 and Fig. 1.

For each element $T \in \mathscr{T}_h$, the elements of B are chosen as follows:

$$B|_T = \begin{pmatrix} b_1 & b_3 \\ b_3 & b_2 \end{pmatrix}, \qquad b_1, b_2, b_3 \in \{-1, 0, 1\}, \qquad \forall T \in \mathscr{T}_h.$$

In other words, we generate a constant perturbation of magnitude δ in each element T. A perturbation generated in this way is clearly an extreme one. It suits our purposes, since we are trying to find a worst case situation that can occur with different diffusion matrices A which belong to the set $\mathscr{D}$.

Let $\mathbb{E}$denote an error indicator computed on the mesh $\mathscr{T}_h$, which generates a set of non-negative numbers associated with elements, i.e.,

$$\mathbb{E}(A, u_h) := \{\mathbb{E}_T\}, \qquad \mathbb{E}_T \geq 0, \qquad \forall T \in \mathscr{T}_h.$$

Its input typically consists of the material data A, and a numerical solution u_h. The output is the vector $\{\mathbb{E}_T\}$, which contains an approximated error value $\mathbb{E}_T$ for each element T.

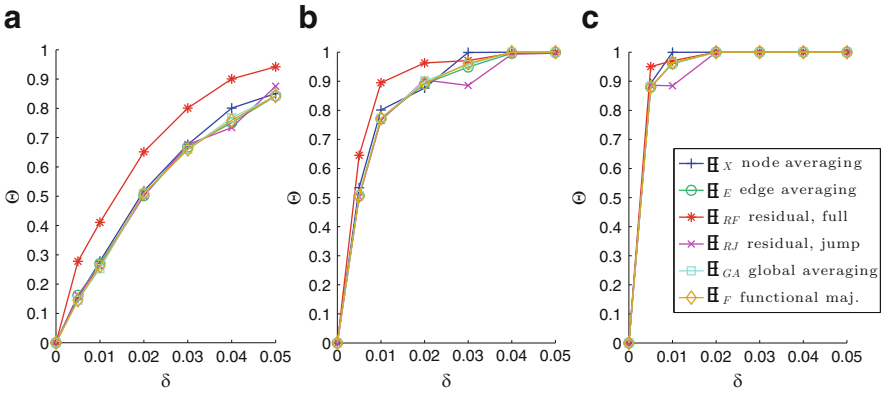


Fig. 1 The values of Θ plotted against the magnitudes of variation δ, for three meshes (example (10)). **(a)** $Nelem = 800$. **(b)** $Nelem = 12800$. **(c)** $Nelem = 115200$

In computational practice, error indicators are used together with a *marker* k that marks elements (or other subdomains) where errors are excessively high. The function k takes as its input the vector produced by an error indicator $\mathbb{E}$, and returns a boolean function indicating by 1 the elements which are to be refined, and by 0 all other elements. The output of $\mathsf{k}(\mathbb{E})$ is essentially the list of those elements T, which contain the majority of the error (according to the indicator used). We refer to the boolean output of k as a *marking*. The marker can, for example, choose to mark some percentage of the elements ("bulk criterion"), or to mark those elements whose indicator value is greater than the average of all the values. In this short note we confine ourselves to the case where k marks a certain amount (N_{ref}) of elements, where the highest values of errors have been indicated.

Our analysis of effects caused by data uncertainty is based on the following method. Let $\mathbb{E}$ be the indicator to test. We select a mesh $\mathscr{T}_h$ and select a certain amount of matrices $A_j = A_0 + \delta B_j$ for some given δ (uncertainty parameter). For each exact solution $u_j = \mathscr{S}(A_j)$, we compute the corresponding approximations u_{jh} on the mesh $\mathscr{T}_h$. Then, for each u_{jh}, we calculate the error indicator $\mathbb{E}_j = \mathbb{E}(A_j, u_{jh})$, and the corresponding markings $\mathsf{k}(\mathbb{E}_j)$.

The difference of two markings is given by the boolean measure

$$\mathrm{diff}(\mathsf{k}, \mathbb{E}_i, \mathbb{E}_j) := 1 - \frac{\sum(\mathsf{k}(\mathbb{E}_i) \wedge \mathsf{k}(\mathbb{E}_j))}{N_{ref}} \in [0, 1],$$

where $\wedge$ is the logical multiplication operator. If $\mathrm{diff}(\mathsf{k}, \mathbb{E}_i, \mathbb{E}_j) = 0$, then small variations of the data do not affect the process of marking. In opposite, if $\mathrm{diff}(\mathsf{k}, \mathbb{E}_i, \mathbb{E}_j)$ is close to one, then the lists of elements selected for refinement by $\mathbb{E}_i$ and $\mathbb{E}_j$ are quite different.

The maximal difference between all markings is given by the quantity

$$\Theta := \max_{i,j}\{\mathrm{diff}(\mathsf{k}, \mathbb{E}_i, \mathbb{E}_j)\},$$

which shows the maximal difference produced by an error indicator with different diffusion matrices from the set $\mathscr{D}$.

From now on, we will denote by u_h an approximation of (1) and (2) calculated with the help of standard linear Courant elements.

We have tested the following six most commonly used error indicators.

Indicators based on averaging. The well known node averaging indicator (see, e.g., [13, 14]) reads

$$\mathbb{E}_{X,T} := \|G_X u_h - A\nabla u_h\|_{A^{-1};T}. \tag{4}$$

and a similar indicator (we call the edge averaging indicator, see, e.g., [11]) reads

$$\mathbb{E}_{E,T} := \|P^p(G_E u_h) - A\nabla u_h\|_{A^{-1};T}. \tag{5}$$

The averaging operators G_X and G_E are defined by the relations

$$G_X u_h(X) = \sum_{T\in\omega_X} \frac{|T|}{|\omega_X|}(A\nabla u_h)|_T \quad \text{and} \quad G_E u_h(E) = \frac{|E|}{\#\omega_E}\sum_{T\in\omega_E}(A\nabla u_h)|_T \cdot n_E,$$

which define the values of $G_X u_h$ and $G_E u_h$ at the node X and edge E, respectively. Then, the averaged function $G_X u_h$ is defined by piecewise affine extension, and $G_E u_h$ by extension with the help of linear Raviart-Thomas elements (see, e.g., [10]). In (5), the operator P is a post-processing operator, which produces more accurate approximations for the exact flux $A\nabla u$ by minimizing the residual $\|f + \operatorname{div} y\|^2_{\omega_E}$ on all subdomains ω_E (see, e.g., [1, 11]). Here $y \in H(\operatorname{div}, \Omega)$ is a vector valued function generated by the averaging operator G_E (we assume that $P^0(G_E u_h) = G_E u_h$).

Residual based indicators. Residual based error indicators form the class of mostly used error indicators (see, e.g., [2, 13]). We consider the standard residual error indicator

$$\mathbb{E}_{RF,T} = \left(h_T^2\|f_T\|_T^2 + \frac{1}{2}\sum_{E\in\mathscr{E}_h(T)\setminus\mathscr{E}_{h,\partial\Omega}} |E|\,\|[n_E\cdot A\nabla u_h]_E\|_E^2\right)^{1/2}, \tag{6}$$

and the indicator containing only jump terms

$$\mathbb{E}_{RJ,T} = \left(\frac{1}{2}\sum_{E\in\mathscr{E}_h(T)\setminus\mathscr{E}_{h,\partial\Omega}} |E|\,\|[n_E\cdot A\nabla u_h]_E\|_E^2\right)^{1/2}. \tag{7}$$

Here f_T denotes the mean value of f on T, i.e., $f_T := \frac{1}{|T|}\int_T f\,dx$.

Global averaging indicator. The global averaging indicator (see, e.g., [3,4]) reads

$$\mathbb{E}_{GA,T} := \|y_{GA} - A\nabla u_h\|_{A^{-1};T}, \tag{8}$$

where y_{GA} is calculated by global minimization of $\|y_{GA} - A\nabla u_h\|^2_{A^{-1};\Omega}$. This minimization procedure results in the problem: find $y_{GA} \in H(\mathrm{div}, \Omega)$ such that

$$\int_\Omega A^{-1} y_{GA} \cdot w \, dx = \int_\Omega \nabla u_h \cdot w \, dx, \qquad \forall w \in H(\mathrm{div}, \Omega).$$

In our tests we used linear Raviart-Thomas finite elements (see, e.g., [10]) in order to find globally averaged indicator on the mesh $\mathscr{T}_h$.

Error indicator generated by the functional type error majorant. The difference between the exact solution u and an approximation u_h is bounded from above by the functional error majorant $M_\oplus$ (see, e.g., [9,11]):

$$\|\nabla(u - u_h)\|^2_{A;\Omega} \le C_1 \|f + \mathrm{div}\, y_F\|^2_\Omega + C_2 \|y_F - A\nabla u_h\|^2_{A^{-1},\Omega} := M_\oplus(A, u_h, y_F),$$

where $C_1 = (1+\alpha) C^2_\Omega \underline{c}^{-1}$ and $C_2 = (1+\alpha^{-1})$. The constant C_Ω is the Friedrich's constant. The above inequality holds for all $y_F \in H(\mathrm{div}, \Omega)$ and $\alpha \in \mathbb{R}_+$. The latter term in the upper bound $M_\oplus$ can be used as an error indicator (see [11] for the mathematical justification of this indicator):

$$\mathbb{E}_{F,T} := \|y_F - A\nabla u_h\|_{A^{-1};T}. \tag{9}$$

The function y_F is calculated by minimization of $M_\oplus$. This minimization procedure results in a problem for $y_F \in H(\mathrm{div}, \Omega)$ and $\alpha \in \mathbb{R}_+$:

$$\int_\Omega \Big(C_1 \mathrm{div}\, y_F \, \mathrm{div}\, w + C_2 A^{-1} y_F \cdot w \Big) dx = \int_\Omega \big(C_2 \nabla u_h \cdot w - C_1 f \, \mathrm{div}\, w \big) dx, \quad \forall w \in H(\mathrm{div}, \Omega).$$

This problem was also solved with the help of linear Raviart-Thomas finite elements.

4 Numerical Results and Conclusions

Approximate solutions of the problem (1) and (2) have been computed using standard Courant type finite element approximations. Indicators (8) and (9) were calculated with the help of linear Raviart-Thomas finite elements. All the problems were calculated on same regular meshes, and systems of linear simultaneous equations were solved by exact methods. In view of this fact, approximate solutions possess Galerkin orthogonality property, and, therefore, the residual error indicator (6) can be used. For the edge averaging indicator (5), we set $p = 5$ (the amount of

times P is applied). All calculations were performed with the MATLAB computing environment on a 64 processor SMP server with 1 TB of RAM.

In total, a mesh contains $N_{elem} := \#\mathscr{T}_h$ elements. Since in this paper we calculate approximations of (1) and (2) using linear Courant elements, the amount of degrees of freedom N_{dof} of an approximation equals the number of nodes $\#\mathscr{N}_h$. We chose to mark 30 % of elements of a mesh to be refined, i.e., $N_{ref} = 0.3 \times N_{elem}$.

We studied how the magnitude of variations δ affects error indicators, and discuss typical results with the paradigm of a simple problem where

$$\Omega = [0,1]^2, \qquad A_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \text{and} \quad f = 2(x_1(1-x_1) + x_2(1-x_2)). \tag{10}$$

The exact solution of this problem is $u_0 = x_1(1-x_1)x_2(1-x_2)$.

Using the procedure explained in Sect. 3, we have tested six different indicators for six different meshes. The results are exposed on Table 1 and Fig. 1. It is worth to outline again, that the actual sensitivity of error indicators with respect to data uncertainty is even higher than in the results reported below (which should be viewed as lower bounds of the true sensitivity).

Table 1 shows how the values of Θ (associated with the indicators (4)–(9)) depend on the amount of elements N_{elem} (or amount of degrees of freedom N_{dof}) and the parameter δ. It is easy to see that sufficiently small values of Θ (which correspond to relatively stable performance of an error indicator) are obtained only for small δ (such as 0.005 or 0.01) and rather moderate amount of elements. If values of δ are not very small (e.g., 0.04) then all indicators may generate quite different markings. We recall that $\Theta = 1$ if indicators computed for different elements of the solution set $\mathscr{D}$ may generate completely opposite markings.

A selection of these numbers are presented on Fig. 1 in a graphical way, which allows us to compare different indicators with each other. We conclude that even in this very simple problem small uncertainties in the matrix coefficients may seriously corrupt the process of error indication. This phenomenon does not depend on a particular error indicator. Actually, it shows that in real life computations error indication procedures (and subsequent mesh refinement) cannot be performed without an adequate analysis of data uncertainty.

Acknowledgements This work was supported by the Finnish foundations Emil Aaltosen säätiö and KAUTE-säätiö, and the COMAS graduate school.

References

1. Anjam, I., Mali, O., Neittaanmäki, P., Repin, S.: A new error indicator for the Poisson problem. In: Mäkinen, R., Neittaanmäki, P., Tuovinen, T., Valpe, K. (eds.), Proceedings of the 10th Finnish Mechanics Days, pp. 324–330 (2009)

2. Babuška, I., Rheinboldt W.C.: Error estimates for adaptive finite element computations. SIAM J. Numer. Anal. **15**, pp. 736–754 (1978)
3. Bartels, S., Carstensen, C.,: Each averaging technique yields reliable a posteriori error control in FEM on unstructured grids. Part II: Higher order FEM. Math. Comput. **71**(239), pp. 971–994 (2002)
4. Carstensen, C., Bartels, S.: Each averaging technique yields reliable a posteriori error control in FEM on unstructured grids. Part I: Low order conforming, nonconforming, and mixed FEM. Math. Comput. **71**(239), pp. 945–969 (2002)
5. Hlaváček, I., Chleboun, J., Babuška, I.: Uncertain input data problems and the worst scenario method. Elsevier, Amsterdam (2004)
6. Mali, O., Repin, S.: Estimates of accuracy limit for elliptic boundary value problems with uncertain data. Adv. Math. Sci. Appl. **19**(2), pp. 525–537 (2009)
7. Mali, O., Repin, S.: Estimates of the indeterminacy set for elliptic boundary value problems with uncertain data. J. Math. Sci. **150**(1), pp. 1869–1874 (2008)
8. Mali, O., Repin, S.: Two-sided estimates of the solution set for the reaction-diffusion problem with uncertain data. Comput. Methods Appl. Sci. **15**, pp. 183–198 (2010)
9. Neittaanmäki, P., Repin, S.: Reliable methods for computer simulation. Error control and a posteriori estimates. Elsevier, Amsterdam (2004)
10. Raviart, P.A., Thomas, J.M.: Primal hybrid finite element methods for 2nd order elliptic equations. Math. Comput. **31**, pp. 391–413 (1977)
11. Repin, S.: A posteriori estimates for partial differential equations. Walter de Gruyter, Berlin (2008)
12. Schuëller, G.I.: A state-of-the-art report on computational stochastic mechanics. Prob. Engrg. Mech. **12**(4), pp. 197–321 (1997)
13. Verfürth, R.: A review of a posteriori error estimation and adaptive mesh-refinement techniques. Wiley and Sons, Teubner-Verlag, New York (1996)
14. Zienkiewicz, O.C., Zhu, J.Z.: A simple error estimator and adaptive procedure for practical engineering analysis. Int. J. Numer. Meth. Engrg. **24**, pp. 337–357 (1987)

A Reduced Basis Method for the Simulation of American Options

B. Haasdonk, J. Salomon, and B. Wohlmuth

Abstract We present a reduced basis method for the simulation of American option pricing. To tackle this model numerically, we formulate the problem in terms of a time dependent variational inequality. Characteristic ingredients are a POD-greedy and an angle-greedy procedure for the construction of the primal and dual reduced spaces. Numerical examples are provided, illustrating the approximation quality and convergence of our approach.

1 Introduction

We consider the problem of American option pricing and refer to [1] and the references therein for an introduction into computational methods for option pricing. While European options can be modelled by a parabolic partial differential equation, American options result in additional inequality constraints. We refer to [9] for a possible numerical treatment by primal-dual finite elements and to [4, 5] for an abstract framework on the theory of constrained variational problems. We are interested in providing a fast numerical algorithm to solve accurately the variational inequality system of an American put option for a large variety of different parameter values such as interest rate, dividend, strike prize and volatility. Reduced basis (RB) methods are an appropriate means for standard parametrized parabolic

B. Haasdonk (✉)
IANS, Universität Stuttgart, Stuttgart, Germany
e-mail: haasdonk@mathematik.uni-stuttgart.de

J. Salomon
CEREMADE, Université Paris-Dauphine, Paris, France
e-mail: salomon@ceremade.dauphine.fr

B. Wohlmuth
M2 – Zentrum Mathematik, Technische Universität München, Münich, Germany
e-mail: wohlmuth@ma.tum.de

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_85,

partial differential equations, cf. [2, 7, 10, 11] and the references therein. These are based on low-dimensional approximation spaces, that are constructed by greedy procedures. Convergence behavior of these procedures are known in some cases [2, 6]. The computational advantage of RB-methods over standard discretization methods is obtained by its possible offline/online decomposition: First, a typically expensive offline-phase involving the computation of the reduced spaces is performed. This phase only needs to be precomputed once. Then, the online phase allows an extremely fast computation of the RB solutions for many new parameters as only low dimensional systems need to be solved. Recently, we adopted the RB methodology to constrained stationary elliptic problems [8], which we extend here to the instationary case. We refer to the recent contribution [3] for a tailored RB approach in option pricing. In contrast to our setting no inequality constraints are taken into account. The main challenge is the construction of a suitable low dimensional approximation of the dual cone required for the approximation of the constraints. In this contribution, we introduce a new greedy strategy based on an angle criteria and show numerical results.

2 American Option Model

An American option is a contract which permits its owner to receive a certain payoff $\psi(S,\tau) \geq 0$ at any time τ between 0 and $T > 0$. The variable T indicates the maturity. Introducing the backward time variable $t := T - \tau$, we can use, e.g., [1] the following non linear model

$$\partial_t P - \frac{1}{2}\sigma^2 s^2 \partial^2_{ss} P - (r-q)s\partial_s P + rP \geq 0, \quad P - \psi \geq 0,$$

$$\left(\partial_t P - \frac{1}{2}\sigma^2 s^2 \partial^2_{ss} P - (r-q)s\partial_s P + rP\right) \cdot \left(P - \psi\right) = 0,$$

where $P = P(s,t)$ is the price of an American put, with $s \in \mathbb{R}_+$ the asset's value, σ is the volatility, r is the interest rate, q is the dividend payment and $\psi = \psi(s,t)$ is the payoff function. The boundary and initial conditions are as follows: $P(s,0) = \psi(s)$, $P(0,t) = K$, $\lim_{s\to+\infty} P(s,t) = 0$, where $K > 0$ is a fixed strike price that satisfies $K = \psi(0,0)$. In what follows, we focus on the case $\psi(s,t) = (K-s)_+$ with $(\cdot)_+ = \max(0,\cdot)$, but our method applies as well to other types of payoff functions. For the implementation, we restrict the values of s to a bounded interval $\Omega := (0, s_f)$, where s_f is large enough to make the assumption $P(s_f,t) = 0$ realistic. Let us also set $\widetilde{P} = P - P_0$, with initial data $P_0(s,t) = K(1 - s/s_f)$, so that $\widetilde{P}$ satisfies homogeneous Dirichlet conditions. Our aim is now to reformulate the last system in a weak form, where our reduced basis method applies. In this view, we introduce the following functional spaces:

$$V := \left\{ v \in L^2(\Omega) | s\partial_s v \in L^2(\Omega), v_{|\partial\Omega} = 0 \right\}, \quad W := V'.$$

The scalar product $\langle \cdot, \cdot \rangle_V$ associated with V is defined by $\langle u, v \rangle_V := \langle s\partial_s u, s\partial_s v \rangle_{L^2(\Omega)} + \langle u, v \rangle_{L^2(\Omega)}$, where $\langle \cdot, \cdot \rangle_{L^2(\Omega)}$ is the usual scalar product on $L^2(\Omega)$. The operators are specified as follows:

$$a(u, v; \mu) = \frac{1}{2}\sigma^2 \langle \partial_s u, \partial_s(s^2 v) \rangle_{L^2(\Omega)} + \langle -(r-q)s\partial_s u + ru, v \rangle_{L^2(\Omega)},$$
$$f(v; \mu) = \langle F, v \rangle_{L^2(\Omega)}, \; g(\eta; \mu) = \langle \widetilde{\psi}, \eta \rangle_W,$$

with $F := -\left(\partial_t P_0 - \frac{1}{2}\sigma^2 s^2 \partial^2_{ss} P_0 - (r-q)s\partial_s P_0 + rP_0\right)$, i.e. $F = K\left(\frac{s}{s_f}q - r\right)$ and $\widetilde{\psi} := \psi - P_0$. For $\eta \in W = V'$, we also define $b(\eta, v) = \eta(v)$. We can now recast our problem in the following weak form, parametrized by $\mu = (K, r, q, \sigma) \in \mathscr{P} \subset \mathbb{R}^4$. We now introduce u as a weak representative of the solution $\widetilde{P}$, as this is the standard notation in reduced basis literature:

$$\langle \partial_t u, v \rangle_{L^2(\Omega)} + a(u, v; \mu) - b(\lambda, v) = f(v; \mu), \qquad v \in V \tag{1}$$
$$b(\eta - \lambda, u) \geq g(\eta - \lambda; \mu), \qquad \eta \in M, \tag{2}$$

where $M \subset W$ is a closed convex cone. Various methods can be considered to solve numerically Eq. (1) and (2). In what follows, we use a θ-scheme for the time discretization. Given $\mu \in \mathscr{P}$, $L \in \mathbb{N}$ and $\Delta t := T/L$, this method corresponds to the following iteration.

Given $0 < n \leq L-1$ and $u^n \in V$, find $u^{n+1} \in V$ and $\lambda^{n+1} \in M$ that satisfy $\forall v \in V, \forall \eta \in M$,

$$\left\langle \frac{u^{n+1} - u^n}{\Delta t}, v \right\rangle_{L^2(\Omega)} + a(\theta u^{n+1} + (1-\theta)u^n, v; \mu) - b(\lambda^{n+1}, v) = f(v; \mu), \tag{3}$$
$$b(\eta - \lambda^{n+1}, u^{n+1}) \geq g(\eta - \lambda^{n+1}; \mu). \tag{4}$$

This recursive definition is initialized with $u^0 := \widetilde{\psi}$. Note that in this scheme, the definition of λ^n is not recursive.

3 Reduced Basis Method

Standard finite element approaches do not exploit the structure of the solution and for a given parameter value, a high dimensional system has to be solved. In what follows, we introduce a specific Galerkin approximation of the solution, based on the reduced basis method and present algorithms to compute the corresponding bases. The principle of the reduced basis method consists in computing parametric solutions in low dimensional subspaces of V and W that are generated with particular solutions of our problem. Let us explain in more detail the corresponding

formulation. For $N \in \mathbb{N}$, consider a finite subset $\mathscr{P}_N := \{\mu_1, \ldots, \mu_N\} \subset \mathscr{P}$ with $\mu_i \neq \mu_j, \ \forall i \neq j$. The reduced spaces V_N and W_N are defined by $V_N := \operatorname{span}\{\psi_1, \ldots, \psi_{N_V}\}$ and $W_N := \operatorname{span}\{\xi_1, \ldots, \xi_{N_W}\}$ where ψ_i and ξ_i are defined from the large set of snapshot solutions $u^n(\mu_i)$ and $\lambda^n(\mu_i)$, $i = 1, \ldots, N$, $n = 0, \ldots, L$. Here $u^n(\mu_i)$ and $\lambda^n(\mu_i)$ denote the solution of Eq. (3) and (4) at the time $t_n := n\Delta t$ for the parameter value $\mu = \mu_i$. The functions ψ_j and ξ_j are suitably selected elements spanning V_N and W_N with $N_V, N_W \leq N(L+1)$ preferably small. Both families $\Psi_N = (\psi_j)_{j=1,\ldots,N_V}$ and $\Xi_N = (\xi_j)_{j=1,\ldots,N_W}$ are supposed to be composed of linearly independent functions, hence are so called reduced bases. Numerical algorithms to build these two sets will be presented in Sect. 4. We define the reduced cone $M_N \subset M$ as

$$M_N = \left\{ \sum_{j=1}^{N_W} \alpha_j \xi_j, \ \alpha_j \geq 0 \right\}.$$

In this setting, the reduced problem reads:

Given $\mu \in \mathscr{P}$, $0 \leq n \leq L-1$, $u_N^n \in V_N$, find $u_N^{n+1} \in V_N$ and $\lambda_N^{n+1} \in M_N$ that satisfy $\forall v_N \in V_N, \forall \eta_N \in M_N$,

$$\left\langle \frac{u_N^{n+1} - u_N^n}{\Delta t}, v_N \right\rangle_{L^2(\Omega)} + a(\theta u_N^{n+1} + (1-\theta)u_N^n, v_N; \mu) - b(\lambda_N^{n+1}, v_N) = f(v_N; \mu), \quad (5)$$

$$b(\eta_N - \lambda_N^{n+1}, u_N^{n+1}) \geq g(\eta_N - \lambda_N^{n+1}; \mu), \quad (6)$$

where the initial value u_N^0 is chosen as the orthogonal projection of u_0 on V_N, i.e.

$$\langle u_N^0 - u_0, v_N \rangle_V = 0, \ \forall v_N \in V_N.$$

4 Reduced Basis Construction

In this section, we present two methods to extract a basis $\Psi_N \subset V$ and $\Xi_N \subset M$ from the snapshots. Both are greedy procedures based on a finite training set $\mathscr{P}_{train} \subset \mathscr{P}$ small enough such that it can be scanned quickly. Given an arbitrary integer N_W, the dual reduced basis $\Xi_N = (\xi_j)_{j=1,\ldots,N_W}$ is built iteratively according to the following algorithm. The goal of the approach is to obtain a reduced cone $M_N \subset M$ capturing as much "volume" as possible.

Algorithm 1. *(Angle-greedy algorithm) Given N_W, $\mathscr{P}_{train} \subset \mathscr{P}$, choose arbitrarily $0 \leq n_1 \leq L$ and $\mu_1 \in \mathscr{P}_{train}$ and do*

1. Set $\Xi_N^1 = \left\{ \frac{\lambda^{n_1}(\mu_1)}{\|\lambda^{n_1}(\mu_1)\|_W} \right\}$, $W_N^1 := \operatorname{span}(\Xi_N^1)$,

2. *For* $k = 1, \dots, N_W - 1$, *do*

   (a) *Find* $(n_{k+1}, \mu_{k+1}) := \operatorname{argmax}_{n=0,\dots,L,\ \mu\in\mathscr{P}_{train}} \left(\measuredangle \left(\lambda^n(\mu), W_N^k \right) \right)$,

   (b) *Set* $\xi_{k+1} := \frac{\lambda^{n_{k+1}}(\mu_{k+1})}{\|\lambda^{n_{k+1}}(\mu_{k+1})\|_W}$,

   (c) *Define* $\Xi_N^{k+1} = \Xi_N^k \cup \{\xi_{k+1}\}$, $W_N^{k+1} := \operatorname{span}(\Xi_N^{k+1})$,

3. *Define* $\Xi_N := \Xi_N^{N_W}$, $W_N := \operatorname{span}(\Xi_N)$.

Here we have used the notation $\measuredangle(v, S)$ to denote the angle between a vector v and a linear space $S \subset W$, which is simply obtained via the orthogonal projection Π_S from W on S by

$$\measuredangle(v, S) = \arccos \frac{||\Pi_S v||_W}{||v||_W}, \quad v \in W.$$

We apply the POD-greedy algorithm [7] to design the primal reduced basis Ψ_N. This procedure is standard in RB-methods for evolution problems. In RB-methods, frequently *weak* greedy procedures are used, which make beneficial use of rapidly computable error estimators and allow to handle large sets $\mathscr{P}_{train}$ [2]. However, as our analysis does not yet provide a-posteriori error estimators, we use the true projection errors as error indicators. This corresponds to the so called *strong* greedy procedure [2, 6].

Algorithm 2. *(POD-greedy algorithm) Given* $\widetilde{N}_V > 0$, $\mathscr{P}_{train} \subset \mathscr{P}$, *choose arbitrarily* $\mu_1 \in \mathscr{P}_{train}$,

1. *Set* $\widetilde{\Psi}_N^1 = \left\{ \frac{u^0(\mu_1)}{\|u^0(\mu_1)\|_V} \right\}$, $\widetilde{V}_N^1 := \operatorname{span}(\widetilde{\Psi}_N^1)$,
2. *For* $k = 1, \dots, \widetilde{N}_V - 1$, *do*

   (a) *Define* $\mu_{k+1} := \operatorname{argmax}_{\mu\in\mathscr{P}_{train}} \left(\sum_{n=0}^L \|u^n(\mu) - \Pi_{\widetilde{V}_N^k}(u^n(\mu))\|_V^2 \right)$,

   (b) *Define* $\widetilde{\psi}_{k+1} := POD_1 \left(u^n(\mu_{k+1}) - \Pi_{\widetilde{V}_N^k}(u^n(\mu_{k+1})) \right)_{n=0,\dots,L}$,

   (c) *Define* $\widetilde{\Psi}_N^{k+1} := \widetilde{\Psi}_N^k \cup \left\{ \widetilde{\psi}_{k+1} \right\}$,

3. *Define* $\widetilde{\Psi}_N := \widetilde{\Psi}_N^{\widetilde{N}_V}$, $\widetilde{V}_N := \operatorname{span}\widetilde{\Psi}_N$.

Here, we have denoted by $\Pi_{\widetilde{V}_N^k}$ the orthogonal projection on $\widetilde{V}_N^k$ with respect to $\langle \cdot, \cdot \rangle_V$, and by POD_1 the routine that extracts from a family of vectors the first Proper Orthogonal Decomposition (POD) mode that can be obtained via the best approximation property

$$POD_1 \left(v^n \right)_{n=0,\dots,L} := \arg \min_{||z||_V = 1} \sum_{n=0}^L ||v^n - \langle v^n, z \rangle_V z||_V^2.$$

In this definition V is spanned by v^n, $n = 0, \dots, L$. A convergence analysis of the POD-greedy procedure is provided in [6]. Note that Algorithm 2 always returns

an orthonormal basis. This is even the case if a parameter value $\mu \in \mathscr{P}_{train}$ is selected more than once. We point out that our System (5) and (6) has a saddle point structure. Thus taking spanΨ_N as reduced basis for the primal variable might result in an ill posed problem. To guarantee the inf-sup stability of our approach, we follow an idea introduced in [10] for the Stokes problem, see also [8] for variational inequalities. It consists in the enrichment $\Psi_N := \widetilde{\Psi}_N^{\widetilde{N}_V} \cup \left(B\xi_i\right)_{i=1,\ldots,N_W}$, where $B\xi_i$ is the solution of $b(\xi_i, v) = \langle B\xi_i, v\rangle_V$, for $v \in V$. We conclude with the final reduced space $V_N := \text{span}\Psi_N$ of dimension $N_V := \dim V_N$. By construction we have $\widetilde{N}_V \leq N_V \leq \widetilde{N}_V + N_W$.

5 Numerical Results

In this section, we present some numerical results obtained on the American Option model. We start with a description of the numerical values and methods we use. In order to compute snapshots, we use a standard finite element method for the space discretization and the θ-scheme presented in Sect. 2 for the time-discretization. The time domain $[0, T] = [0, 1]$ is discretized with a uniform mesh of step size $\Delta t := T/L$, $L = 20$. The θ-scheme is used with $\theta = 1/2$, i.e. we apply a Crank-Nicolson method. The space domain $\Omega = (0, s_f) = (0, 300)$ is discretized with a uniform mesh of step size $\Delta s := s_f/S$, $S = 101$. For the function space, we use standard conforming nodal first order finite elements. For the sake of simplicity, we keep the notation V for the discrete high dimensional space and define it by $V := \{v \in H_0^1(\Omega) | v_{|[s_m, s_{m+1}]} \in P_1, m = 0, \ldots, S-1\}$ of dimension $H_V = H := S - 2 = 99$ with $s_m := m\Delta s$. We associate the basis function $\phi_i \in V$ with its Lagrange node $s_i \in \Omega$, i.e., $\phi_i(s_j) = \delta_{ij}, i, j = 1, \ldots, H$. The discretization of the Lagrange multipliers is performed using a dual finite element basis χ_j of $W := V'$ having the same support as ϕ_j, so that $b(\phi_i, \chi_j) = \delta_{ij}, i, j = 1, \ldots, H_W = H$. The cone M is defined by: $M = \left\{\sum_{i=1}^{H_W} \eta_i \chi_i,\ \eta_i \geq 0\right\}$. To build the basis, we consider a subset $\mathscr{P}_{train}$ of $\mathscr{P}$ that is composed of $N = 16$ values chosen randomly in the set

$$\begin{aligned}\mathscr{P} = &[(1 - \tfrac{\varepsilon}{2})K_0, (1 + \tfrac{\varepsilon}{2})K_0] \times [(1 - \tfrac{\varepsilon}{2})r_0, (1 + \tfrac{\varepsilon}{2})r_0] \\ &\times [(1 - \tfrac{\varepsilon}{2})q_0, (1 + \tfrac{\varepsilon}{2})q_0] \times [(1 - \tfrac{\varepsilon}{2})\sigma_0, (1 + \tfrac{\varepsilon}{2})\sigma_0].\end{aligned}$$

with the numerical values $\varepsilon = 0.1$, $K_0 = 100$, $r_0 = 0.05$, $q_0 = 0.0015$, $\sigma_0 = 0.5$. To define the basis Ψ_N and the convex set Ξ_N, we use Algorithm 2 combined with the enlargement by the supremizers and Algorithm 1. The eight first vectors of Ψ_N, Ξ_N and the supremizers are represented in Fig. 1. We simulate two trajectories corresponding to the values $(\widetilde{N}_V, N_W) = (8, 8)$ and $(\widetilde{N}_V, N_W) = (16, 16)$ respectively. The corresponding bases Ψ_N are of size $N_V = 16$ and $N_V = 32$ respectively. We chose randomly a parameter vector μ corresponding to the values $K = 106.882366$, $r = 0.048470$, $d = 0.007679$, $\sigma = 0.418561$ in $\mathscr{P}$. Some steps of the simulation

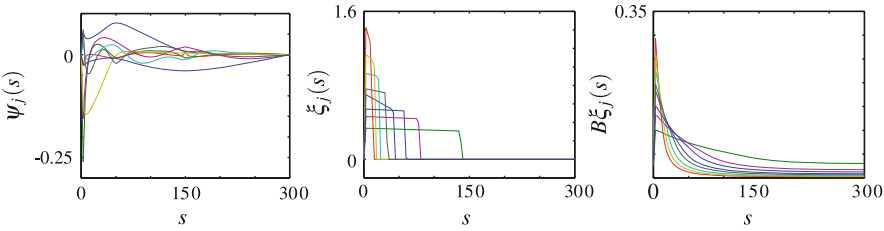


Fig. 1 Eight first vectors of the reduced basis Ψ_N, Ξ_N and the corresponding supremizers

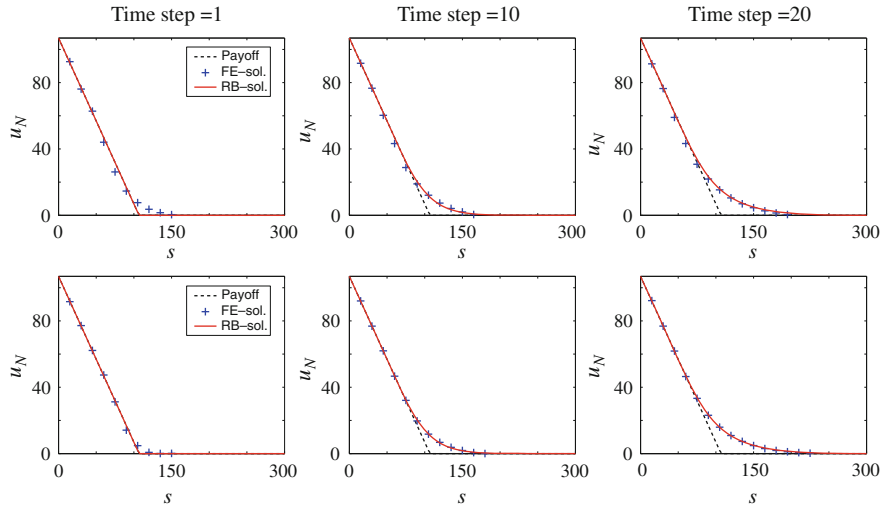


Fig. 2 Finite element approximation *(solid red line)* and Reduced basis approximation *(blue +)* at time steps $t/\Delta t = 1$, $t/\Delta t = 10$ and $t/\Delta t = T/\Delta t = 20$. The payoff function ψ is represented with the *black dashed line*. The reduced bases that are used have been generated by $(\widetilde{N}_V, N_W) = (8, 8)$ *(plots on the top)* or $(\widetilde{N}_V, N_W) = (16, 16)$ *(plots on the bottom)*

are represented in Fig. 2, the top and lower row refer to the smaller and larger reduced spaces, respectively. We clearly see the improvement in the approximation by increasing the reduced dimensions. In order to evaluate the efficiency of the greedy algorithms proposed in Sect. 4, we plot the evolution of the quantities

$$\varepsilon_N^u := \max_{\mu \in \mathscr{P}_{train}} \sqrt{\sum_{n=0}^{L} \|u^n(\mu) - \Pi_{V_N^k}(u^n(\mu))\|_V^2}, \quad \varepsilon_N^\lambda := \max_{\substack{n=0,\dots,L,\\ \mu \in \mathscr{P}_{train}}} \left(\measuredangle \left(\lambda^n(\mu), W_N^k \right) \right)$$

during their iterations. The results are plotted in the first two diagrams in Fig. 3. We observe an excellent exponential convergence of the approximation measures.

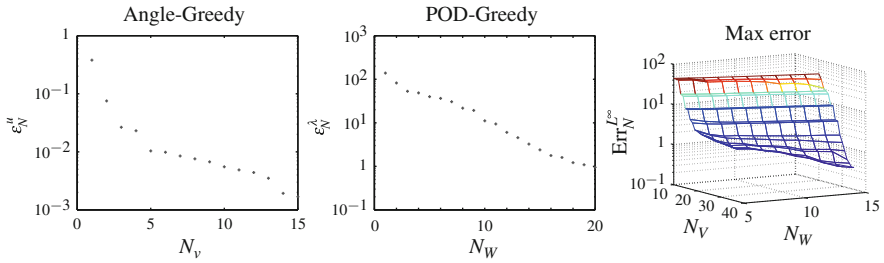


Fig. 3 Values of ε_N^u and ε_N^λ during the iterations of the greedy Algorithms 1 *(left)* and 2 *(middle)*. *Right*: Values of $\mathrm{Err}_N^{L^\infty}$ with respect to N_V and N_W

As final experiment, we address the generalization ability of the RB-model to parameters outside the training set. We consider $\mathscr{P}_{test} \subset \mathscr{P}$, a random set of $N_{test} = 10$ parameter vectors and estimate, for a given $\mu \in \mathscr{P}$, the efficiency of our method through these quantities:

$$\mathrm{err}_N(\mu) = \sqrt{\Delta t \sum_{n=0}^{L} \|u^n(\mu) - u_N^n(\mu)\|_V^2}, \quad \mathrm{Err}_N^{L^\infty} = \max_{\mu \in \mathscr{P}_{test}} \left(\mathrm{err}_N(\mu)\right).$$

Note that $\mathrm{err}_N(\mu)$ actually depends on Ψ_N; for the sake of simplicity, we have omitted this dependence in the notation. As a test, we evaluate the influence of the parameters $\widetilde{N}_V, N_W$ determining the sizes of the bases Ψ_N and Ξ_N on $\mathrm{Err}_N^{L^\infty}$. The results are plotted in the right diagram of Fig. 3. In our example we numerically obtain $N_V = \widetilde{N}_V + N_W$ in all cases, indicating, that the primal snapshots and supremizers are linearly independent. We observe a reasonable good error decay when simultaneously increasing $\widetilde{N}_V$ and N_W, indicating that the reduced method is working well. We also note that in our case, the size of the dual basis has a limited impact on the results.

References

1. Y. Achdou and O. Pironneau. *Computational methods for option pricing*. Frontiers in applied mathematics. Society for Industrial and Applied Mathematics, 2005.
2. Annalisa Buffa, Yvon Maday, Anthony T. Patera, Christophe Prud'homme, and Gabriel Turinici. A priori convergence of the greedy algorithm for the parametrized reduced basis. *To be published ESAIM-Math. Model. Numer. Anal.*, 2011.
3. Rama Cont, Nicolas Lantos, and Olivier Pironneau. A reduced basis for option pricing. *SIAM Journal on Financial Mathematics*, 2:287–316, 2011.
4. C. Geiger and C. Kanzow. *Theorie und Numerik restringierter Optimierungsaufgaben*. Springer-Lehrbuch Masterclass. Springer, 2002.
5. R. Glowinski. *Numerical Methods for Nonlinear Variational Problems*. Scientific Computation. Springer, 2008.

6. Bernard Haasdonk. Convergence rates of the POD-greedy method. jun 2011. SimTech Preprint 2011-23, University of Stuttgart, submitted.
7. Bernard Haasdonk and Mario Ohlberger. Reduced basis method for finite volume approximations of parametrized linear evolution equations. *M2AN, Math. Model. Numer. Anal*, 42(2):277–302, 2008.
8. Bernard Haasdonk, Julien Salomon, and Barbara Wohlmuth. A reduced basis method for parametrized variational inequalities. *submitted*, 2011. SimTech Preprint 2011-16, University of Stuttgart.
9. C. Hager, S. Hüeber, and B. Wohlmuth. Numerical techniques for the valuation of basket options and its greeks. *J. Comput. Fin.*, 13(4):1–31, 2010.
10. G. Rozza. *Shape design by optimal flow control and reduced basis techniques: applications to bypass configurations in haemodynamics*. PhD thesis, EPFL, Lausanne, 2005.
11. K. Veroy, C. Prud'homme, D. V. Rovas, and A. T. Patera. A posteriori error bounds for reduced-basis approximation of parametrized noncoercive and nonlinear elliptic partial differential equations. In *In Proceedings of 16th AIAA computational fluid dynamics conference*, 2003. Paper 2003-3847.

SAVU: A Statistical Approach for Uncertain Data in Dynamics of Axially Moving Materials

J. Jeronen

Abstract In physics and engineering problems, model input is never exact. The effect of small uncertainties on the solution is thus an important question. In this study, a direct statistical-visual approach to approximate the solution set is investigated in the context of axially moving materials. The multidimensional probability distribution for the input uncertainties is assumed known. It is considered as a deterministic object, which is then mapped through the model. The resulting probability density of the model output is visualized. The proposed system consists of three non-trivial parts, which are briefly discussed: a multidimensional sampler, a density estimator, and a high dynamic range (HDR) plotter. Dynamic range compression is achieved via tone mapping techniques from HDR photography. This allows a contrast-preserving representation of the HDR data on a regular computer display or on paper. The model itself is treated as a black box; hence, the same approach can be used for investigating e.g. floating-point rounding errors or approximation error instead of uncertain data.

1 Introduction

In physics and engineering problems, model input is never exact. The effect of small input uncertainties on the model output is thus an important problem both qualitatively and quantitatively. To obtain reliable analysis results, it is desirable to find out how stable the predictions of a given model are with respect to small perturbations in input, and quantitatively how large the expected range of output is.

In problems with uncertain data, instead of a single solution, one obtains a solution set corresponding to the set of admissible inputs. One makes the distinction

J. Jeronen (✉)
Department of Mathematical Information Technology, University of Jyväskylä, Mattilanniemi 2 (Agora), 40014, Jyväskylä, Finland
e-mail: juha.jeronen@jyu.fi

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_86,

between *uncertainty analysis* and *sensitivity analysis*. Uncertainty analysis aims at determining, in model output, the uncertainty which derives from uncertainty in model inputs. Sensitivity analysis aims at determining the contribution of individual uncertain inputs to the uncertainty in the output (see e.g. [6, 12]).

Many different methods exist for uncertainty analysis. In the context of partial differential equation (PDE) models, there are for example a priori error estimates, functional a posteriori error estimates (e.g. [14, 16]), and statistical methods. Some methods require only error margins for the input, while others utilize information of the shape of the input probability distribution.

Which method is the best depends on the use case. The functional a posteriori approach offers deep analysis of certain standard model classes (e.g. elliptic or parabolic PDE). The statistical approach, on the other hand, is extremely flexible. Different input probability distributions can be easily specified, and many different kinds of models can be analyzed quickly.

This study concentrates only on uncertainty analysis, and approaches it from the statistical viewpoint. We assume a known N-dimensional probability distribution for the input uncertainties. The distribution is considered as a deterministic object, which is then mapped through the model. Similar ideas have been explored by e.g. J. C. Helton [10–12]; statistical methods for uncertainty analysis date back at least to the 1970s. See e.g. the review [6], which collects both deterministic and statistical methods. Statistical methods are reviewed also in [12].

Treating the model as a black box, we run an ensemble simulation to approximate the resulting solution set directly. Original to this study, the result is then visualized as a "smoke cloud", with shade representing the probability density.

Running the actual ensemble simulation, being an embarrassingly parallel step, is in a sense trivial. However, it can only be done on a discrete sample of input, which has to be produced first. Also the approximation of the continuum probability density, based on the corresponding sample of model output, is nontrivial.

Multidimensional sampling and density estimation are standard subjects of statistics research; see e.g. [2, 15, 17, 20]; and [5, 7, 8, 19]; respectively. The high dynamic range plotter in this study is original, based on tone mapping techniques from HDR photography [9, 21]. Using tone mapping for plotting mathematical functions can be seen as a natural extension of the ideas of Park and Montag [18], who investigated the use of tone mapping for representation of data from astronomical and medical imaging (at wavelengths other than visible light).

2 Overview of SAVU

The basic idea is sketched in Fig. 1. We limit our considerations to the case where the admissible input set is a cartesian product of one-dimensional ranges. Each variable comes with a one-dimensional probability distribution, which are combined to form the joint N-dimensional input distribution. After equal-probability binning, the admissible input set forms an N-dimensional hypercube. The input variables

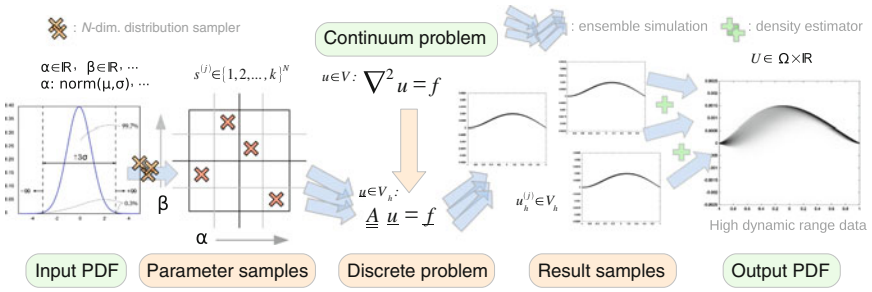


Fig. 1 Overview of SAVU (Sample-based Analysis and Visualization of Uncertainty): from an input probability distribution function (PDF) to output

are assumed statistically independent (uncorrelated). This assumption is reasonable, if we are interested in analyzing the effects of measurement error in the inputs, since we can assume the errors are independent. Strictly speaking, the variables themselves should be independent from the viewpoint of the physics described by the model, but for small relative measurement error (small input range) this should not matter.

The sampling of the input probability distribution function (PDF) is performed using a Latin hypercube sampler (LHS) with a sudoku constraint (see Fig. 1; in addition to only one sample per hyperplane, we require only one sample in each N-dimensional "large box"). The sudoku constraint guarantees a coarse N-dimensional stratification on top of the one-dimensional stratification of the classical LHS of [15]. Many other methods exist, but "sudoku LHS" runs in linear time with linear memory usage (in terms of target sample size), is simple to implement,[1] and stratifies slightly better (in terms of lower pairwise correlation of variables in output) than classical LHS. The improvement is the most noticeable in two dimensions (because then we have stratification in both one and $N = 2$ dimensions), but a slight improvement can be seen for higher dimensions, too, at least up to $N = 4$. Of course, a more advanced sampling method such as one of [2, 17, 20] could also be used.

Stratification of the sample—ideally in all subsets of the set of input coordinate axes—is essential if we do not wish to assume anything of the form of the response function of the model being analyzed. Note especially that, as is well-known, changing just one input variable at a time (one-at-a-time, OAT) should be avoided, because the OAT strategy only scans an N-dimensional plus sign shaped region in the input space. Unless the response function being analyzed is of the additive form

[1] MATLAB implementations of the various components of SAVU are available at http://yousource.it.jyu.fi/savu/codes/
Link referred 13th January 2012.

$f(x_1, x_2, \ldots, x_N) = f_1(x_1) + f_2(x_2) + \cdots + f_N(x_N)$, OAT cannot guarantee to capture the dominant behaviour for the whole hypercube of admissible inputs.[2]

After the ensemble simulation for the chosen input samples is completed, the resulting probability density is estimated based on the obtained output samples. The density estimate is constructed by kernel density estimation, which is a nonparametric way of approximating the density of a set of points in one or more dimensions. It is generally agreed that the choice of the actual kernel function does not matter much; the Gaussian is often used. A much more difficult issue—especially in two or more dimensions—is finding a suitable bandwidth matrix for the kernel. For discussion and suggestions, see e.g. [5, 7, 8, 19].

In [5], a fast Fourier-based density estimator is described. The method is extremely fast in both one and two dimensions. In the present study, this method was used. However, the kernel bandwidth was chosen manually to slightly smooth out the density estimate, because it was found that the automatic bandwidth estimation included in the algorithm produces bandwidths which look overly sharp, rendering the discrete nature of the output sample clearly visible.

Finally, the resulting probability density data is approximately logarithmically distributed (in the histogram sense) and comes with a high dynamic range. Dynamic ranges of well over 100 dB are common based on performed tests. If the shade of gray is used to encode function value, regular computer displays with 8 bits per colour channel can only handle $20 \cdot \log_{10} 255/1 \approx 48.1$ dB of dynamic range.

Due to the logarithmic distribution of the data, linear scaling causes most of the data to fall into the first few bins, rendering a blank image. Logarithmic scaling (effectively plotting the exponent), on the other hand, displays everything but eliminates visual contrast. The standard image processing method of histogram equalization exaggerates small contrasts [21]. See Fig. 2.

Hence, in the present study, *tone mapping* techniques from HDR photography were used to achieve dynamic range compression for contrast-preserving plotting of mathematical functions with high dynamic range. This produces a data-adaptive colour scale which is neither linear nor logarithmic. The motivation for tone mapping in the photographic context is that the human visual system is sensitive to contrasts, but not to absolute brightnesses (see e.g. [21]). Thus, large contrasts can be compressed.

In order to achieve a globally meaningful colour scale (keeping the plot quantitative), a *global* tone mapping method must be chosen. Global methods assign always the same colour to the same function value (light intensity) in the same picture, while local methods may deviate from this.

The contrast expansion limiting algorithm of [21] was found to produce excellent results, while being very simple to implement. A one-dimensional trivial modification of the gradient attenuation method of [9], operating in the logarithmic histogram domain, was also tested. With a white-to-black grayscale (white zero, black maximum), the results were found consistently slightly darker than in the

[2]Consider $f(x_1, x_2) = \varepsilon x_1 + \varepsilon x_2 + C x_1 x_2$, where $C/\varepsilon \gg 1$, in a square centered at the origin.

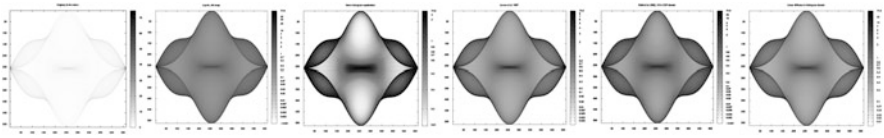


Fig. 2 Dynamic range compression for HDR plotting. *From left to right*: Linear; log; histogram equalization; contrast expansion limiter of [21]; 1D log histogram gradient attenuation based on [9]; linear diffusion in log histogram domain

case of [21]; the method of [9] allocates a relatively larger portion of the total colour scale to the low end (small function values), leaving relatively fewer colours to the high end. Finally, as an original method, linear diffusion (the heat equation) in the logarithmic histogram domain was tested. This method smooths out the logarithmic histogram, while retaining (a smoothed version of) the most massive peaks. All three methods produce acceptable results, but judging them visually (see Fig. 2), [21] wins by a small margin. As it is also the simplest to implement, [21] is the recommended choice.

3 Application Examples

In this section, we display some elementary application examples. Three physical examples are considered: steady state diffusion in one dimension, the time behaviour of a one degree of freedom harmonic oscillator, and travelling (d'Alembert) waves in a finite string with reflecting ends.

Note that for vibration problems, of which the last two are examples, the behaviour of the displacement itself is not very interesting, except possibly for realtime control purposes (where predictions in the extreme short term are enough). As is well-known, for a non-damped oscillator, any small disturbance in the frequency will cause two initially close trajectories to eventually diverge. For linear vibration problems, it is generally more interesting to look at the eigenfrequencies.

In this section, we will simply apply the method to the displacement variable, thus visually confirming this well-known observation. In the next section, discussing the axially travelling panel, we will look also at eigenfrequencies.

In Fig. 3, left, one-dimensional steady state diffusion is shown. The equation is $-cu_{xx} = f$, with c uncertain and $f = 0.85 = \text{const}$. The boundary conditions are zero Dirichlet. The value of $c = c_0 + X$, where $c_0 = 0.5$ and X is a random variable with a Gaussian distribution truncated at $\pm 3\sigma$. The standard deviation was chosen as 5 % of the reference value, i.e. $\sigma = 0.05c_0$. As this leads to the total uncertainty range 6σ being 30 % of c_0, we see that the one-dimensional diffusion problem is very stable with respect to measurement errors in the diffusion coefficient, as expected.

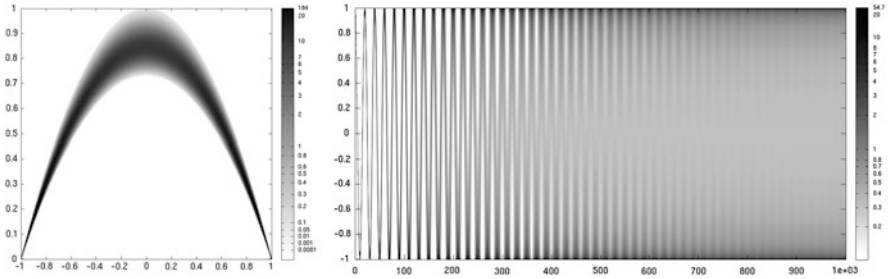


Fig. 3 Examples. *Shade* indicates probability density. *Left*: One-dimensional steady state diffusion, uncertain diffusion coefficient. *Right*: One degree of freedom harmonic oscillator, uncertain frequency. *Horizontal axis*: time. *Vertical axis*: oscillator position

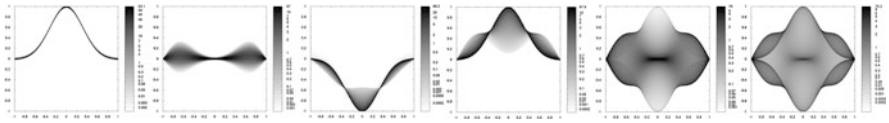


Fig. 4 Travelling wave in a finite string with reflecting ends. Snapshots at different times. *Shade* indicates probability density. Initial pulse at *left*, final state at *right*

Figure 3, right, displays the behaviour of a harmonic oscillator with one degree of freedom, at an uncertain frequency. We have taken $u(t) = \cos(\pi ct)$, with c uncertain (same distribution as above, now with $c_0 = 0.1$ and $\sigma = 0.01c_0$). The position of the oscillator spreads out, and eventually fills the whole interval $[-1, 1]$. This is as expected for a vibration problem. Note that even in this extremely simple example, a quick calculation confirms that the time required for Poincaré recurrence is so astronomical that it will not be seen in practice for any reasonable sample size; once the solutions have spread out, they will stay spread out for any reasonable t.

A travelling wave in a finite string with reflecting ends is illustrated in Fig. 4, with the initial pulse shown in the first picture. The behaviour is similar to the oscillator; as enough time passes, all admissible positions of the periodic solution become filled. In this example, it is seen more clearly that the local probability density is proportional to the time any given solution spends in a particular region.

4 The Axially Travelling Panel Submerged in Ideal Fluid

We consider an axially travelling panel submerged in ideal fluid (potential flow), with an optional axial free-stream component. Details can be found in our studies [1, 13]. *Panel* is understood as a plate in the limit of cylindrical deformation (the *flat panel* of aeroelasticity; see e.g. [3]).

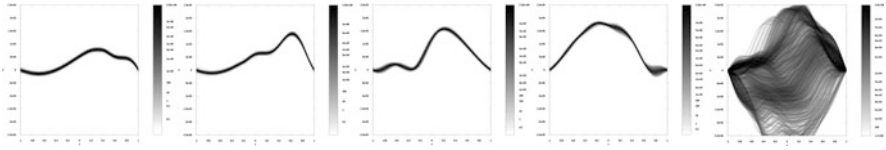


Fig. 5 Axially travelling panel submerged in ideal fluid. Snapshots taken at different times; the final image much later than the others. *Shade* indicates probability density

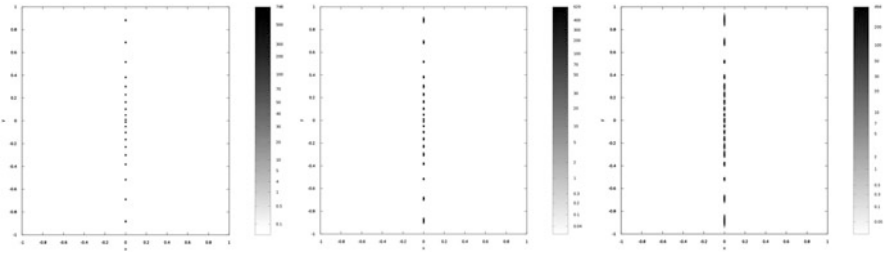


Fig. 6 Eigenfrequencies of axially travelling panel submerged in ideal fluid. Uncertain panel velocity. Relative uncertainty is, from *left* to *right*, 1, 5 and 10 %

We will look at both the out-of-plane displacement of the panel, and at its complex eigenfrequencies (the stability exponent of Bolotin, [4]).

First, in Fig. 5, we have snapshots of the time behaviour of the displacement, with four simultaneously uncertain parameters. The input is thus a four-dimensional hypercube. The uncertain parameters are the axial panel velocity V_0, applied axial tension T, the Young modulus of the panel E, and the mass per unit area of the panel m. A Gaussian distribution truncated at $\pm 3\sigma$ is used for each parameter, representing typical measurement error. The standard deviation σ is taken (arbitrarily) as 1 % of the reference value independently for each parameter. The governing equation and reference values for the parameters can be found in [13, p. 100, 155].

Figure 6 represents the eigenfrequencies of a moving panel. Validity flagging (see [13]) has not been performed. For this particular model with the particular Galerkin basis used, only the lowest n_0 modes (in absolute value), of the $2 \cdot n_0$ computed, represent physically meaningful solutions.

For the eigenfrequency problem, we cannot use the density estimator of [5], because all the eigenvalues are located on the imaginary axis even in the presence of uncertainties. Thus, there is a very sharp edge in the data along the horizontal direction at the imaginary axis, triggering severe ringing artifacts in Fourier-based methods. Instead, to produce the eigenvalue plots presented here, a simplistic, direct space domain Gaussian kernel density estimator was implemented as a backup.

5 Conclusion

A statistical-visual method for uncertainty analysis was outlined. Multidimensional sampling, density estimation, and visualization of high dynamic range mathematical functions were briefly discussed.

Some simple application examples, and selected simulations from a real research case on axially moving materials, were presented. Although only linear examples were considered for the sake of simplicity, it should be emphasized that the method is especially well suited for nonlinear models due to the black-box nature. The same method can also be used for investigating e.g. floating-point rounding errors or approximation error instead of uncertain data.

Finally, the method is useful when one wishes to try out a variety of different models for a physics or engineering problem, not necessarily sharing the same mathematical properties.

References

1. Banichuk, N. V., Jeronen, J., Neittaanmäki, P. J., Tuovinen, T.: Dynamic behaviour of an axially moving plate undergoing small cylindrical deformation submerged in axially flowing ideal fluid. J. Fluids Struct. **27**, 986–1005 (2011) doi:10.1016/j.jfluidstructs.2011.07.004
2. Beachkofski, B. K., Grandhi, R. V.: Improved distributed hypercube sampling. In: 43rd AIAA structures, structural dynamics, and materials conference. AIAA-2002-1274, Denver, CO (2002)
3. Bisplinghoff, R. L., Ashley, H.: Principles of Aeroelasticity. 2nd edition. Dover, New York (1975)
4. Bolotin, V. V.: Nonconservative Problems of the Theory of Elastic Stability. Pergamon Press, New York (1963)
5. Botev, Z. I., Grotowski, J. F., Kroese, D. P.: Kernel Density Estimation via Diffusion. Ann. Stat. **38**(5), 2916–2957 (2010)
6. Cacuci, D. G., Ionescu-Bujor, M.: A Comparative Review of Sensitivity and Uncertainty Analysis of Large-Scale Systems. Nucl. Sci. Eng. **147**(3), 189–203 (Part I: Deterministic methods), 204–217 (Part II: Statistical Methods) (2004)
7. Chacón, J. E., Duong, T.: Multivariate plug-in bandwidth selection with unconstrained pilot bandwidth matrices. Test. **19**, 375–398 (2010)
8. Duong, T., Hazelton, M. L.: Plug-in bandwidth selectors for bivariate kernel density estimation. J. Nonparam. Stat. **15**, 17–30 (2003)
9. Fattal, R., Lischinski, D., Werman, M.: Gradient Domain High Dynamic Range Compression. In: Proceedings of ACM SIGGRAPH (2002)
10. Helton, J. C., Davis, F. J.: Illustration of sampling-based methods for uncertainty and sensitivity analysis. Risk Anal. **22**(3), 591–622 (2002)
11. Helton, J. C., Davis, F. J.: Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. Technical report, Sandia National Laboratories, SAND2001-0417 (2002)
12. Helton, J. C., Johnson, J. D., Sallaberry, C. J, Storlie, C. B.: Survey of sampling-based methods for uncertainty and sensitivity analysis. Technical report, Sandia National Laboratories, SAND2006-2901 (2006)

13. Jeronen, J.: On the mechanical stability and out-of-plane dynamics of a travelling panel submerged in axially flowing ideal fluid: a study into paper production in mathematical terms. PhD thesis, University of Jyväskylä, Jyväskylä (2011) http://julkaisut.jyu.fi/?id=978-951-39-4596-1
14. Mali, O., Repin, S.: Estimates of accuracy limit for elliptic boundary value problems with uncertain data. Adv. Math. Sci. Appl. **19**, 525–537 (2009)
15. McKay, M. D., Beckman, R. J., Conover, W. J.: A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. Technometrics **21**(2), 239–245 (1979)
16. Neittaanmäki, P., Repin, S. I.: Reliable methods for computer simulation: error control and a posteriori estimates. Elsevier (2004)
17. Owen, A. B.: Orthogonal Arrays for Computer Experiments, Integration and Visualization. Stat. Sin. **2**, 439–452 (1992)
18. Park, S. H., Montag, E. D.: Evaluating Tone Mapping Algorithms for Rendering Non-Pictorial (Scientific) High-Dynamic-Range Images. J. Vis. Commun. Image Represent. **18**(5), 415–428 (2007)
19. Sheather, S. J., Jones, M. C.: A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation. J. Royal Stat. Soc., Series B (Methodological) **53**(3), 683–690 (1991)
20. Tang, B.: Orthogonal Array-Based Latin Hypercubes. J. Amer. Stat. Assoc. **88**(424), 1392–1397 (1993)
21. Ward–Larson, G., Rushmeier, H., Piatko, C.: Visibility Matching Tone Reproduction Operator for High Dynamic Range Scenes. IEEE Trans. Vis. Comp. Graph. **3**, 291–306 (1997)

On Singularity of Fisher Information Matrix for Stochastic Processes Under High Frequency Sampling*

R. Kawai

Abstract We discuss the singularity of the Fisher information arising from statistical inference for continuous-time stochastic processes of practical interest, such as asset price dynamics in finance and individual animal movement in biology, under high frequency discrete sampling schemes. Singularity seems to be caused by the scale parameter and the selfsimilarity index, while there exists a different type of singularity resulting from some redundancy of parameters in the short time framework. We derive the speed of convergence of the Fisher information to singularity for some instances and show that the convergence to singularity may be delayed through a wise expansion of the total observation window.

1 Introduction

For a continuous-time stochastic process model based on high frequency data, one of the most fundamental, yet in no way obvious, issues is estimation of the characterizing parameters involved. Suppose that we are given n observations of the stochastic process $\{X_t(\theta) : t \geq 0\}$ in $\mathbb{R}$, indexed by the parameter $\theta \in \mathbb{R}^d$, at equidistant time points $\{t_{n,k}\}_{k=1,\ldots,n}$ such that $t_{n,k} := k\Delta_n$, with the mesh Δ_n of the grid decreasing to zero;

$$t_{n,k+1} - t_{n,k} = \Delta_n \downarrow 0,$$

as $n \uparrow +\infty$. The number n is usually taken very large, hence one is interested in asymptotic properties as n increases to infinity. It is essential that our interest

*This work was carried out largely while the author was based at University of Leicester, UK

R. Kawai (✉)
School of Mathematics and Statistics, University of Sydney, NSW 2006, Australia
e-mail: reiichiro.kawai@maths.usyd.edu.au

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_87,

for estimation lies in the parameter θ of the continuous-time stochastic process $\{X_t(\theta) : t \geq 0\}$, not in the discrete-time random walk $\{X_{t_{n,k}}(\theta)\}_{k=1,\ldots,n}$. High frequency sampling has attracted much attention due to increasingly availability of high resolution data, for example, of asset price dynamics in finance and individual animal movement in biology. The high frequency sampling, which is technically meant by $\Delta_n \downarrow 0$, reflects the best possible realistic experiment environment. In the other words, strictly speaking, observation over a whole interval is never possible even with recent high technology; for example, although high resolution video recording may look like providing a continuous movement, it is still discrete at an extraordinarily high frequency.

The local asymptotic normality property is a vital concept in asymptotically optimal statistical analyses. In short, the local asymptotic normality property for a differentiable statistical model for the parameter θ to be estimated is defined through the weak convergence of the likelihood ratio to the Gaussian shift experiment; for each $h \in \mathbb{R}^d$,

$$\frac{d\mathbb{P}_{\theta+R_n(\theta)h}}{d\mathbb{P}_\theta}\Big|_{\mathscr{F}_n} \xrightarrow{\mathscr{L}} \exp\left[\langle h, Z(\theta)\rangle - \frac{1}{2}\langle h, \mathscr{I}(\theta)h\rangle\right], \tag{1}$$

under $\mathbb{P}_\theta$, where $\mathbb{P}_\theta|_{\mathscr{F}_n}$ is a probability measure associated with θ restricted to the filtration $\mathscr{F}_n$, $\{R_n(\theta)\}_{n\in\mathbb{N}}$ is a sequence of diagonal matrices in $\mathbb{R}^{d\times d}$ whose diagonal entries are positive and tend to zero, $\mathscr{I}(\theta)$ is a non-negative definite deterministic matrix in $\mathbb{R}^{d\times d}$, called the Fisher information matrix, and $Z(\theta) \sim \mathscr{N}(0, \mathscr{I}(\theta))$ under $\mathbb{P}_\theta$. If the above weak convergence holds, then we say that *the local asymptotic normality property holds at point θ with the rate $R_n(\theta)$ and the Fisher information matrix $\mathscr{I}(\theta)$*. If the local asymptotic normality property holds with non-singular $\mathscr{I}(\theta)$, then a unbiased estimator $\{\widehat{\theta}_n\}_{n\in\mathbb{N}}$ of θ is said to be asymptotically efficient in a neighborhood of θ if

$$R_n(\theta)^{-1}\left(\widehat{\theta}_n - \theta\right) \xrightarrow{\mathscr{L}} \mathscr{N}\left(0, \mathscr{I}(\theta)^{-1}\right),$$

under $\mathbb{P}_\theta$, that is, such estimators achieve asymptotically the Cramér-Rao lower bound $\mathscr{I}(\theta)^{-1}$ for the estimator variance. (We refer the reader to [1, 3, 8, 11] for thorough details.)

2 Singularity of Fisher Information

We discuss some examples of stochastic processes of practical interest, whose Fisher information is singular under high frequency sampling. To illustrate some aspects of the present work, let us begin with the simplest non-singular case of the Brownian motion.

2.1 Illustrative Example: Brownian Motion

Let $\{B_t : t \geq 0\}$ be a standard Brownian motion in $\mathbb{R}$ and define

$$X_t := t\gamma + \sigma B_t, \quad t \geq 0.$$

Then, it is a standard fact that the local asymptotic normality property holds with

$$\theta = \begin{bmatrix} \gamma \\ \sigma^2 \end{bmatrix}, \ R_n(\theta) = \begin{bmatrix} \frac{1}{\sqrt{n\Delta_n}} & 0 \\ 0 & \frac{1}{\sqrt{n}} \end{bmatrix}, \ \mathscr{I}(\theta) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}.$$

This result can be proved without difficulty, as this experiment for each $n \in \mathbb{N}$ is simply about iid Gaussian sample thanks to the independence and the stationarity of increments of the Brownian motion, whose probability density function is available in closed form. Clearly, the Fisher information matrix $\mathscr{I}(\theta)$ is invertible. In order to consistently estimate the drift γ, we require the increasing total length of the observed interval, that is, $t_{n,n} = n\Delta_n \uparrow +\infty$. Note however that we may still efficiently estimate the variance σ^2 alone without forcing $n\Delta_n \uparrow +\infty$. This is so due to the diagonality of $\mathscr{I}(\theta)$; any asymptotically efficient joint estimation of (γ, σ^2) is asymptotically mutually independent and hence γ can be left as a unknown nuisance parameter. This fact is consistent with the intuition. On the one hand, as the global behavior of sample paths can be described by the drift parameter γ, the observation window is required to expand for the drift γ. Otherwise, if one might mistakenly focus on a fixed interval with very negatively skewed path, then estimation would indicate a negative drift although the true γ were positive. The variance σ^2, on the other hand, governs a local behavior of sample paths, which may be estimated from a fixed interval, without looking at the global tendency of sample paths.

Moreover, based upon the above fact, let us point out the interesting fact that high frequency sampling in the limit does not corresponds to continuous sampling. This is counter-intuitive, for example, compared to the definition of the Riemann integral. In fact, it is known by the Girsanov theorem that the measures $\mathbb{P}_{\theta+R_n(\theta)h}|_{\mathscr{F}_n}$ here are mutually equivalent if and only if the parameter σ is in common. Otherwise, the likelihood ratio $(d\mathbb{P}_{\theta+R_n(\theta)h}/d\mathbb{P}_\theta)|_{\mathscr{F}_n}$ in (1), or equivalently the left hand side of (1), would not be well defined. Indeed, if the whole sample path were observable, then the true value of σ could be computed exactly, that is, with no estimation error at all, by computing the quadratic variation of the sample path. Hence, through the analysis of high frequency sampling schemes, we would be interested in the phenomenon with an extremely small $\Delta_n > 0$, rather than with its limit $\Delta_n = 0$.

2.2 Fractional Brownian Motion

Let $\{B_t^H : t \geq 0\}$ be a standard fractional Brownian motion in $\mathbb{R}$ with the Hurst parameter $H \in (0, 1)$, that is,

$$\mathbb{E}_\theta\left[B_t^H B_s^H\right] = \frac{1}{2}\left(t^{2H} + s^{2H} - (t-s)^{2H}\right), \quad 0 \le s \le t,$$

and define

$$X_t := t\gamma + \sigma B_t^H, \quad t \ge 0.$$

For simplicity, we fix $\gamma = 0$, and then it is strictly selfsimilar with index H. We exclude the case $H = 1$, since then $B_t^1 = tB_1^1$, $a.s.$, which is nothing but a straight line. Then, the local asymptotic normality property holds [6] with

$$\theta = \begin{bmatrix} \sigma \\ H \end{bmatrix}, \ R_n(\theta) = \begin{bmatrix} \frac{1}{\sqrt{n}} & 0 \\ 0 & \frac{1}{\sqrt{n}|\ln \Delta_n|} \end{bmatrix}, \ \mathscr{I}(\theta) = \begin{bmatrix} \frac{2}{\sigma^2} & -\frac{2}{\sigma} \\ -\frac{2}{\sigma} & 2 \end{bmatrix}. \tag{2}$$

Unless $H = 1/2$, this experiment is no longer about iid Gaussian sample, while the increments are still identically distributed Gaussian and the likelihood function is available in closed form. As a matter of course, if $H = 1/2$ is a priori known, then the model reduces immediately to the Brownian motion.

Observe first that the scale σ and the selfsimilarity H cause the singularity of the Fisher information matrix. Also surprisingly, the information of H is independent of H. In principle, the parameter H has much to do with the regularity of nowhere differentiable sample paths; trajectories are Hölder continuous of any order strictly less than H, that is, sample paths are rougher with smaller H. We may better think of this experiment as estimation of the above covariance structure, as such regularity of sample paths is simply difficult to estimate based on discrete observations. Interestingly, it is known [7] that this independence of H does not occur under the low frequency sampling scheme $\Delta_n \equiv \Delta > 0$.

2.3 Stable Lévy Process

Let $\{L_t^{(\alpha)} : t \ge 0\}$ be a symmetric stable process satisfying

$$\mathbb{E}_\theta\left[e^{iyL_t^{(\alpha)}}\right] = \exp\left[-t|y|^\alpha\right], \quad y \in \mathbb{R}, \tag{3}$$

and define

$$X_t := t\gamma + \sigma L_t^{(\alpha)}, \quad t \ge 0.$$

Like the fractional Brownian motion, if $\gamma = 0$, then it is strictly selfsimilar with index α. Then, the local asymptotic normality property holds [10] with

$$\theta = \begin{bmatrix} \gamma \\ \sigma \\ \alpha \end{bmatrix}, \ R_n(\theta) = \begin{bmatrix} \frac{1}{\sqrt{n}\Delta_n^{1/\alpha-1}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{n}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{n}|\ln \Delta_n|} \end{bmatrix}, \ \mathscr{I}(\theta) = \begin{bmatrix} \frac{C_1(\alpha)}{\sigma^2} & 0 & 0 \\ 0 & \frac{C_2(\alpha)}{\sigma^2} & \frac{C_2(\alpha)}{\sigma\alpha^2} \\ 0 & \frac{C_2(\alpha)}{\sigma\alpha^2} & \frac{C_2(\alpha)}{\alpha^4} \end{bmatrix},$$

where $C_1(\alpha)$ and $C_2(\alpha)$ are positive constants depending only on α. (The constant $C_2(\alpha)$ will be defined later.) The Fisher information is singular due to the scale σ and the selfsimilarity α. We have seen a similar phenomenon due to (σ, H) in the case of the fractional Brownian motion in Sect. 2.2 as well.

2.4 Meixner Lévy Process

We next give an example of a different type, namely, singularity not due to the scale and the selfsimilarity. The Meixner Lévy process has been recognized as a successful class of Lévy processes for the purpose of practical modelling, such as mathematical finance and possibly turbulence, as well as of sufficient theoretical interest. Precisely, it is a Lévy process with the marginal density function at time $t > 0$ given in closed form

$$f_t(y;\theta) = \frac{(2\cos(\beta/2))^{2t\delta}}{2\pi\alpha\Gamma(2t\delta)} \exp\left[\frac{\beta}{\alpha}(y-\mu)\right] \left|\Gamma\left(t\delta + i\frac{y-\mu}{\alpha}\right)\right|^2, \quad y \in \mathbb{R},$$

where $\alpha > 0$ indicates the tail heaviness, $\beta \in (-\pi, +\pi)$ the skewness, $\delta > 0$ the time scale, and $\mu \in \mathbb{R}$ the location. It behaves like a Cauchy (namely, 1-stable) Lévy process over a very short period of time. More precisely, it holds that as $h \downarrow 0$,

$$\left\{ \frac{1}{h\alpha\delta}\left(X_{ht} - ht\mu\right) : t \geq 0 \right\} \stackrel{\mathscr{L}}{\to} \left\{ L_t^{(1)} : t \geq 0 \right\}. \tag{4}$$

Hence, under high frequency sampling, roughly speaking, we are observing iid Cauchy sample. It is proved [5] that the local asymptotic normality property holds with

$$\theta = \begin{bmatrix} \alpha \\ \beta \\ \delta \\ \mu \end{bmatrix}, \ R_n(\theta) = \begin{bmatrix} \frac{1}{\sqrt{n}} & 0 & 0 & 0 \\ 0 & \frac{1}{\sqrt{n\Delta_n}} & 0 & 0 \\ 0 & 0 & \frac{1}{\sqrt{n}} & 0 \\ 0 & 0 & 0 & \frac{1}{\sqrt{n}} \end{bmatrix}, \ \mathscr{I}(\theta) = \begin{bmatrix} \frac{1}{2\alpha^2} & 0 & \frac{1}{2\alpha\delta} & 0 \\ 0 & \frac{\delta}{2\cos^2(\beta/2)} & 0 & 0 \\ \frac{1}{2\alpha\delta} & 0 & \frac{1}{2\delta^2} & 0 \\ 0 & 0 & 0 & \frac{1}{2\alpha^2\delta^2} \end{bmatrix},$$

where the Fisher information matrix is obviously singular due to (α, δ). Note that the (short-time) selfsimilarity index is known to be 1 due to the Cauchy convergence (4), unlike in the unknown selfsimilarity index cases, such as the fractional Brownian

motion (Sect. 2.2) and the stable Lévy process (Sect. 2.3). In the meantime, the convergence (4) indicates that the product $\alpha\delta$ acts as a scale parameter in the short time framework. Both α and δ admit the same optimal rate $1/\sqrt{n}$. We may conclude that in the short time framework, either one of α or δ is redundant.

There are some examples of non-singularity for Lévy processes, such as the normal inverse Gaussian Lévy process [4], the inverse Gaussian process and the gamma process [9]. In particular, the normal inverse Gaussian process has the exactly same set of four parameters $(\alpha, \beta, \delta, \mu)$ and enjoys a short-time Cauchy convergence similar to the one (4) of the Meixner Lévy process, while the parameter δ alone acts as the short-time scaling parameter in the Cauchy convergence, unlike $\alpha\delta$ in (4) for the Meixner Lévy process. Also, in the case of the normal inverse Gaussian process, the parameters α and δ admit different optimal rates $1/\sqrt{n\Delta_n}$ and $1/\sqrt{n}$, respectively. Although the domain of β depends on α for the normal inverse Gaussian processes, these two parameters do not cause singularity, while there is no dependency in parameter domains in the Meixner Lévy process. This acts as a concrete counterexample against the intuition that dependency among parameters causes singularity. On the whole, it seems very difficult to make a judgement at first glance of the particular setting in its probability density function whether singularity will come out as an issue.

3 Speed of Convergence to Singularity

We have seen some concrete examples of singular Fisher information under high frequency sampling. There are mainly two conventional conclusions drawn from such singularity. One is that if a parameter causing singularity is a priori known (for instance, either σ or α in the case of the fractional Brownian motion), then asymptotic normality is guaranteed in the limit as the inverse of the resulting (partial) Fisher information matrix is well defined. The other is that joint estimation should be handled with care with an extremely small time stepsize, as the estimator normality is then nearly broken down. It is known that this issue may be addressed by giving up the optimal rate to gain normality of joint estimation [10], or by giving up the estimation variance at the Cramér-Rao lower bound where rate efficiency is retained.

In this section, we take a further closer look at the singularity phenomenon on the basis of the speed of convergence to singularity and extract some useful information for designing the experiment setting. To this end, the Fisher information matrix is required based upon n observations, not in the limit $n \uparrow +\infty$. In the both models, we set the drift parameter aside γ as it has no effect on the singularity in the limit.

In the case of the fractional Brownian motion of Sect. 2.2, recall from (2) that the stepsize Δ_n has only to be decreasing without additional requirements. In this case, the total observation window $n\Delta_n \equiv T > 0$ can even be fixed through $\Delta_n = T/n$ to extract sufficient information for both σ and H (although the existence of both causes singularity).

Now, could an increasing total observation window $n\Delta_n \uparrow +\infty$ be of any help? If yes, in what sense? Those are the questions we are going to investigate. Consider the Fisher information matrix $\mathscr{I}_n^{(\sigma,H)}(\theta)$ for (σ, H) with n observations for the fractional Brownian motion, which can be derived in a similar manner to the proof of [6, Lemma 3.3] as

$$\mathscr{I}_n^{(\sigma,H)}(\theta) := \begin{bmatrix} \frac{2}{\sigma^2} & -\frac{2}{\sigma} - \frac{\mathrm{tr}[B_n(H)]}{n|\ln\Delta_n|\sigma} \\ -\frac{2}{\sigma} - \frac{\mathrm{tr}[B_n(H)]}{n|\ln\Delta_n|\sigma} & 2 - \frac{2\mathrm{tr}[B_n(H)]}{n|\ln\Delta_n|} + \frac{\mathrm{tr}[B_n(H)^2]}{2n|\ln\Delta_n|^2} \end{bmatrix}$$
$$\to \begin{bmatrix} \frac{2}{\sigma^2} & -\frac{2}{\sigma} \\ -\frac{2}{\sigma} & 2 \end{bmatrix} =: \mathscr{I}^{(\sigma,H)}(\theta),$$

where $B_n(H) := (T_n(H)^{1/2})^\top \partial_H (T_n(H))^{-1} T_n(H)^{1/2}$, and $T_n(H)$ is the symmetric nonsingular Toeplitz matrix in $\mathbb{R}^{n\times n}$ defined by $(T_n(H))_{k_1,k_2} = (|k_1 - k_2 + 1|^{2H} - 2|k_1 - k_2|^{2H} + |k_1 - k_2 - 1|^{2H})/2$, for $k_1, k_2 = 1, \ldots, n$. Then, for each $n \in \mathbb{N}$, the determinant is given in closed form by

$$\det\left(\mathscr{I}_n^{(\sigma,H)}(\theta)\right) = \frac{1}{\sigma^2|\ln\Delta_n|^2}\left[\frac{\mathrm{tr}[B_n(H)^2]}{n} - \left(\frac{\mathrm{tr}[B_n(H)]}{n}\right)^2\right]. \tag{5}$$

It is known [2] that both $\mathrm{tr}[B_n(H)^2]$ and $\mathrm{tr}[B_n(H)]$ above behave at most like $O(n)$ at infinity. Hence, the determinant (5) explodes at the rate of $|\ln\Delta_n|^{-2}$, or possibly faster if the difference $(\mathrm{tr}[B_n(H)^2]/n - (\mathrm{tr}[B_n(H)]/n)^2)$ tends to zero as well. For example, with a straightforward choice of $\Delta_n = n^{-1/c}$ for some $c \geq 1$, this rate is $|\ln\Delta_n|^{-2} = c^2|\ln(n)|^{-2}$, indicating that a larger c would (slightly) help. If the total observation window can be expanded faster with $\Delta_n = (\ln(n))^{-1}$, then the convergence of the Fisher information matrix to singularity can be delayed further to the rate $|\ln\Delta_n|^{-2} = |\ln(\ln(n))|^{-2}$. On the whole, an increasing total observation window, if at all possible, indeed helps delay the convergence to singularity.

In the case of the symmetric stable Lévy process of Sect. 2.3, we can derive the (partial) Fisher information matrix $\mathscr{I}_n^{(\sigma,\alpha)}(\theta)$ for (σ, α) with n observations as

$$\mathscr{I}_n^{(\sigma,\alpha)}(\theta) := \begin{bmatrix} \mathscr{I}_n^{11}(\theta) & \mathscr{I}_n^{12}(\theta) \\ \mathscr{I}_n^{12}(\theta) & \frac{C_2(\alpha)}{\sigma^2} \end{bmatrix} \to \begin{bmatrix} \frac{C_2(\alpha)}{\alpha^4} & \frac{C_2(\alpha)}{\alpha^2\sigma} \\ \frac{C_2(\alpha)}{\alpha^2\sigma} & \frac{C_2(\alpha)}{\sigma^2} \end{bmatrix} =: \mathscr{I}^{(\sigma,\alpha)}(\theta)$$

where

$$\mathscr{I}_n^{11}(\theta) = \frac{C_2(\alpha)}{\alpha^4} - \frac{2C_3(\alpha)}{\alpha^2\ln\Delta_n} + \frac{C_4(\alpha)}{|\ln\Delta_n|^2}, \quad \mathscr{I}_n^{12}(\theta) = \frac{C_2(\alpha)}{\alpha^2\sigma} + \frac{C_3(\alpha)}{\sigma|\ln\Delta_n|},$$
$$C_2(\alpha) := \int_{\mathbb{R}} \frac{(f_\alpha(x) + x f'_\alpha(x))^2}{f_\alpha(x)} dx,$$

$$C_3(\alpha) := \int_{\mathbb{R}} \frac{f_\alpha(x) + x f'_\alpha(x)}{f_\alpha(x)} \frac{\int_{\mathbb{R}} |y|^\alpha \ln|y| e^{-iyx-|y|^\alpha} dy}{2\pi f_\alpha(x)} f_\alpha(x) dx,$$

$$C_4(\alpha) := \int_{\mathbb{R}} \left(\frac{\int_{\mathbb{R}} |y|^\alpha \ln|y| e^{-iyx-|y|^\alpha} dy}{2\pi f_\alpha(x)} \right)^2 f_\alpha(x) dx,$$

with f_α being the smooth probability density function of $\mathscr{L}(L_1^{(\alpha)})$ defined in (3). Then, for each $n \in \mathbb{N}$, the determinant is available in closed form by

$$\det\left(\mathscr{I}_n^{(\alpha,\sigma)}(\theta)\right) = \frac{1}{\sigma^2 |\ln \Delta_n|^2} \left[C_1(\alpha) C_3(\alpha) - C_2(\alpha)^2 \right] (\neq 0), \tag{6}$$

indicating the explosion at the rate of $|\ln \Delta_n|^{-2}$. We can prove that the determinant above is not zero for each n, which indicates, with the help of the Cauchy-Schwarz inequality, that with a fixed stepsize $\Delta_n \equiv \Delta > 0$, the scale σ and the selfsimilarity α do not cause singularity here.

Recall from Sect. 2.3 that in the full joint case, the optimal rate for the drift γ requires $\sqrt{n}\Delta_n^{1/\alpha-1} \uparrow +\infty$ to extract sufficient information for its estimation. If $\alpha \in [1, 2)$ (or a priori known γ in $\alpha \in (0, 1)$), then this requirement is automatically satisfied. Thus, just like for the fractional Brownian motion, we wish to expand the total observation window $n\Delta_n \uparrow +\infty$ with the choice of a slow decay, such as $\Delta_n = (\ln(n))^{-1}$, so as to delay the convergence to singularity. If $\alpha \in (0, 1)$ without knowledge of γ, on the other hand, then the stepsize Δ_n is required to decay slower than $n^{\frac{\alpha}{2(\alpha-1)}}$, that is, $\Delta_n = n^{\frac{\alpha}{2(\alpha-1)}+c}$ with some $c \in (0, \alpha/(2-2\alpha))$. Again, an increasing total observation window, if at all possible, would help delay the convergence to singularity. Interestingly, both (5) and (6) converge at the same rate of $|\ln \Delta_n|^{-2}$. This might be the common speed of convergence to singularity caused by scale and selfsimilar.

References

1. Cramér, H., *Mathematical Methods of Statistics*. Princeton University Press, Princeton (1946)
2. Dahlhaus, R., Efficient parameter estimation for self-similar processes, *Ann. Stat.*, **17**(4) 1749–1766 (1989)
3. Jacod, J., Inference for stochastic processes, In: *Handbook of Financial Econometrics*, Aït-Sahalia, Y. and Hansen, L.P. (eds.) Amsterdam, North-Holland (2010)
4. Kawai, R., Masuda, H., Local asymptotic normality for normal inverse Gaussian Levy processes with high frequency sampling, *ESAIM: PS*, doi:10.1051/ps/2011101.
5. Kawai, R., Masuda, H., On the local asymptotic behavior of the likelihood function for Meixner Levy processes under high frequency sampling, *Stat. Probab. Lett.*, **81**(4) 460–469 (2011)
6. Kawai, R., Fisher information for fractional Brownian motion under high frequency discrete sampling, *Commun. Stat. - Theory Methods*, doi:10.1080/03610926.2011.594540.

7. Kawai, R., Petrovskii, S., Multiscale properties of random walk models of animal movement: lessons from statistical inference, *Proc. Roy. Soc. Lond. A*, *Proc. Roy. Soc. Lond. A*, **468**(2141) 1428–1451 (2012)
8. Le Cam, L., Yang, G.L., *Asymptotics in Statistics. Some Basic Concepts. Second Edition*, Springer Series in Statistics. Springer-Verlag, New York (1990)
9. Masuda, H., Notes on estimating inverse-Gaussian and gamma subordinators under high-frequency sampling, *Ann. Inst. Statist. Math.*, **61**, 181–195 (2009)
10. Masuda, H., Joint estimation of discretely observed stable Lévy processes with symmetric Lévy density, *J. Japan Statist. Soc.*, **39**(1) 49–75 (2009)
11. Rao, C.R., *Linear Statistical Inference and Its Applications*, 2nd ed. Wiley, New York (1973)

Hierarchical Model Reduction: Three Different Approaches

S. Perotto and A. Zilio

Abstract We present three different approaches to model, in a computationally cheap way, problems characterized by strong horizontal dynamics, even though in the presence of transverse heterogeneities. The three approaches are based on the hierarchical model reduction setting introduced in Ern et al. (Hierarchical model reduction for advection-diffusion-reaction problems. In: Kunisch K, Of G, Steinbach O (eds) Numerical mathematics and advanced applications. Springer (2008), pp 703–710) and Perotto et al. (Multiscale Model Simul 8(4):1102–1127, 2010).

1 Motivations

We focus on the modeling of engineering applications which exhibit a dominant dynamic (e.g., flows in tubular domains as in haemodynamics or in a channel network as in hydrodynamics, flows through anisotropic porous media). For this modeling, downscaled models, where only the dominant space dependence is considered, are sometimes advisable. Nevertheless, in the presence of significant transverse dynamics, these downscaled models may become uneffective (see, e.g., [2]).

We move consequently to a different approach, known as Hierarchical Model (Hi-Mod) reduction to get a sort of trade-off between accuracy and efficiency

S. Perotto (✉)
MOX, Dipartimento di Matematica "F. Brioschi", Politecnico di Milano, Piazza Leonardo da Vinci 32, I-20133, Milano, Italy
e-mail: simona.perotto@polimi.it

A. Zilio
Dipartimento di Matematica "F. Brioschi", Politecnico di Milano, Piazza Leonardo da Vinci 32, I-20133, Milano, Italy
e-mail: alessandro.zilio@mail.polimi.it

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_88,

[1, 3]. We suitably rewrite the *full problem* as a set of coupled 1D differential problems (i.e., the *reduced model*) associated with the dominant dynamic, while the information along the transverse directions are lumped in the coefficients of the reduced formulation. We focus on a generic second-order elliptic full problem, given by

$$\text{find } u \in V \quad : \quad a(u, v) = \mathscr{F}(v) \quad \forall v \in V, \tag{1}$$

with $V \subseteq H^1(\Omega)$ a Hilbert space, $a(\cdot, \cdot) : V \times V \to \mathbb{R}$ a continuous and coercive bilinear form and $\mathscr{F}(\cdot) : V \to \mathbb{R}$ a continuous linear functional.

In this paper we propose three different techniques for a Hi-Mod reduction. The first two approaches have already been validated with good results (see [1, 3]). The nodewise Hi-Mod reduction represents the novelty of this paper.

2 Hierarchical Model Reduction Techniques

We fix the basic ingredients to perform a Hi-Mod reduction [1, 3]. We first introduce a constraint on the computational domain. We assume $\Omega = \bigcup_{x \in \Omega_{1D}} \{x\} \times \gamma_x$, i.e., Ω coincides with a *fiber bundle*, where $\Omega_{1D} = (x_0, x_1)$ is the supporting fiber (parallel to the dominant dynamics) while γ_x is the transverse fiber at x (parallel to the secondary transverse dynamics). In particular, we focus on 2D domains.

Then, for any $x \in \Omega_{1D}$, we introduce the map $\psi_x : \gamma_x \to \widehat{\gamma}$ between the generic fiber γ_x and the reference fiber $\widehat{\gamma}$, so that the physical domain Ω is mapped into the reference domain $\widehat{\Omega} = \Omega_{1D} \times \widehat{\gamma}$ via the map $\Psi : \Omega \to \widehat{\Omega}$, given by $\Psi(\mathbf{z}) = \widehat{\mathbf{z}}$, where $\mathbf{z} = (x, y)$, $\widehat{\mathbf{z}} = (\widehat{x}, \widehat{y})$ with $\widehat{x} = x$ and $\widehat{y} = \psi_x(y)$. We assume that ψ_x is a C^1-diffeomorphism for all $x \in \Omega_{1D}$ and that Ψ is differentiable with respect to $\mathbf{z}$. A standard choice for ψ_x is an affine map.

The fiber structure on Ω is at the basis of all the Hi-Mod reduction techniques below. The common idea is to differently tackle the dependence of the full solution on the dominant and on the transverse directions. The former is spanned via a standard (1D) finite element basis. The latter are expanded into a modal basis $\{\varphi_k\}_{k \in \mathbb{N}^+}$ of functions in $H^1(\widehat{\gamma})$, orthonormal with respect to the $L^2(\widehat{\gamma})$-scalar product and compatible with the boundary conditions along the horizontal sides of Ω.

2.1 Uniform Hi-Mod Reduction

This approach resorts to a global one-dimensional space $V_{1D} \subseteq H^1(\Omega_{1D})$ to describe the solution along the fiber Ω_{1D} as well as to the same number of modal functions along the transverse directions [1, 3]. In particular, the functions in V_{1D} take into account the boundary conditions on the vertical sides of Ω.

The discrete uniform Hi-Mod reduced formulation for (1) reads: for a certain modal index $m \in \mathbb{N}^+$,

$$\text{find } u_m^h \in V_m^h \quad : \quad a(u_m^h, v_m^h) = \mathscr{F}(v_m^h) \quad \forall v_m^h \in V_m^h, \tag{2}$$

where the discrete reduced space

$$V_m^h = \left\{ v_m^h(x,y) = \sum_{k=1}^{m} \widetilde{v}_k^h(x)\, \varphi_k(\psi_x(y)), \text{ with } \widetilde{v}_k^h \in V_{1D}^h,\ x \in \Omega_{1D},\ y \in \gamma_x \right\} \tag{3}$$

establishes an actual *hierarchy* of reduced models marked by the modal index m, i.e., by the different level of detail in describing the transverse dynamics of the full problem. Space $V_{1D}^h \subset V_{1D}$ is a finite element space associated with a subdivision $\mathscr{T}_h$ of Ω_{1D}, with $\dim(V_{1D}^h) = N_h < +\infty$. A standard density assumption is made on V_{1D}^h. Then, suitable hypotheses of conformity and of spectral approximability guarantee the inclusion $V_m^h \subset V$ as well as the well-posedness of the reduced formulation (2).

If we replace in (2) the reduced solution with the corresponding discrete modal representation ($u_m^h(\mathbf{z}) = \sum_{k=1}^m \widetilde{u}_k^h(x)\, \varphi_k(\psi_x(y))$) and choose $v_m^h = \vartheta_i \varphi_j$, with ϑ_i the generic finite element basis function, we are led to solve

$$\sum_{k=1}^{m} a(\widetilde{u}_k^h \varphi_k, \vartheta_i \varphi_j) = \mathscr{F}(\vartheta_i \varphi_j) \quad j = 1, \ldots, m,\ i = 1, \ldots, N_h \tag{4}$$

i.e., a set of coupled 1D problems instead of the full 2D problem. From an algebraic viewpoint, (4) coincides with a linear system with an $m \times m$ block matrix, where each block is an $N_h \times N_h$ matrix exhibiting the sparsity of the finite element space.

An appropriate choice of the modal index m in (3) is certainly the most critical issue of the uniform Hi-Mod reduction. This choice can be driven, e.g., by an a priori knowledge of the phenomenon at hand. In [3] a "trial and error" approach is suggested: we move from the computationally cheapest choice for m ($m = 1$) and then we gradually increase such a value. We stop when the addition of the successive modal function does not significantly improve the accuracy of the reduced solution. This choice may become really uneffective when strongly localized transverse dynamics are present. In such a case a large number of modal functions is required on the whole Ω, even though it would be strictly necessary only on the portion of Ω where the strong dynamics occur.

2.2 *Piecewise Hi-Mod Reduction*

To overcome the intrinsic limit of a uniform Hi-Mod reduction, we move in [3] to a new formulation, where a different number of modes is employed in different

parts of Ω: essentially, large values of m are used where the transverse dynamics are relevant, small values where the dynamics are less important. In particular, we resort to a domain decomposition approach to glue the models associated with a different number of modes: the reduced problem is thus split and iteratively solved on subdomains of Ω. The modal index m becomes therefore a piecewise constant vector: this justifies the name of this approach.

Following [4], the discrete piecewise Hi-Mod reduced formulation for (1) reads: for a certain modal multi-index $\mathbf{m} \in [\mathbb{N}^+]^s$,

$$\text{find } u_{\mathbf{m}}^{b,h} \in V_{\mathbf{m}}^{b,h} \quad : \quad a_\Omega(u_{\mathbf{m}}^{b,h}, v_{\mathbf{m}}^{b,h}) = \mathscr{F}_\Omega(v_{\mathbf{m}}^{b,h}) \quad \forall v_{\mathbf{m}}^{b,h} \in V_{\mathbf{m}}^{b,h}, \tag{5}$$

with $a_\Omega(u_{\mathbf{m}}^{b,h}, v_{\mathbf{m}}^{b,h}) = \sum_{i=1}^s a_i(u_{\mathbf{m}}^{b,h}|_{\Omega_i}, v_{\mathbf{m}}^{b,h}|_{\Omega_i})$, $\mathscr{F}_\Omega(v_{\mathbf{m}}^{b,h}) = \sum_{i=1}^s \mathscr{F}_i(v_{\mathbf{m}}^{b,h}|_{\Omega_i})$ where $a_i(\cdot,\cdot)$ and $\mathscr{F}_i(\cdot)$ are the restrictions of the bilinear and linear forms in (1) to the s subdomain Ω_i of Ω, such that $\overline{\Omega} = \cup_{i=1}^s \overline{\Omega}_i$. The modal multi-index $\mathbf{m} = \{m_i\}_{i=1}^s$ collects the number of modes employed on each Ω_i. The discrete reduced space $V_{\mathbf{m}}^{b,h}$ is defined by

$$V_{\mathbf{m}}^{b,h} = \Big\{ v_{\mathbf{m}}^{b,h} \in L^2(\Omega) \ : \ v_{\mathbf{m}}^{b,h}|_{\Omega_i} = \sum_{k=1}^{m_i} \widetilde{v}_k^{i,h}|_{\Omega_{1D,i}}(x)\, \varphi_k(\psi_x(y)) \in H^1(\Omega_i)$$

$$\forall i = 1, \dots, s, \ \text{with } \widetilde{v}_k^{i,h} \in V_{1D}^{b,h} \text{and s.t., } \forall k = 1, \dots, m_\perp^j \text{ with } j = 1, \dots, s-1, \tag{6}$$

$$\int_\gamma \big[v_{\mathbf{m}}^{b,h}|_{\Omega_{j+1}}(\sigma_j, \psi_{\sigma_j}^{-1}(\widehat{y})) - v_{\mathbf{m}}^{b,h}|_{\Omega_j}(\sigma_j, \psi_{\sigma_j}^{-1}(\widehat{y})) \big] \varphi_k(\widehat{y})\, d\widehat{y} = 0 \Big\},$$

with $m_\perp^j = \min(m_j, m_{j+1})$, $\Omega_{1D,i} = \Omega_{1D} \cap \Omega_i$, $\sigma_j = \overline{\Omega}_j \cap \overline{\Omega}_{j+1}$. Space $V_{1D}^{b,h}$ is a suitable discrete space associated with the finite element partition $\mathscr{T}_h$: it represents a subset of the one-dimensional broken Sobolev space $H^1(\Omega_{1D}, \mathscr{T}_{\Omega_{1D}})$ depending on the partition $\mathscr{T}_{\Omega_{1D}} = \{\Omega_{1D,i}\}_{i=1}^s$ of the supporting fiber Ω_{1D}. Likewise, the space $V_{\mathbf{m}}^{b,h}$ is a subset of the two-dimensional broken Sobolev space $H^1(\Omega, \mathscr{T}_\Omega)$ associated with the partition $\mathscr{T}_\Omega = \{\Omega_i\}_{i=1}^s$ of Ω.

Notice that the integral condition in (6) weakly enforces the continuity of the solution in correspondence with the minimum number of modes employed on the whole Ω. This does not guarantee a priori the conformity of the reduced solution $u_{\mathbf{m}}^{b,h}$ in (5). Different strategies can be adopted to impose this interface condition: in [4] we resort to an iterative substructuring Dirichlet/Neumann method (with relaxation).

From a computational viewpoint, at each iteration of the Dirichlet/Neumann scheme, we apply, separately, a uniform Hi-Mod reduction on the subdomains Ω_i. This leads to solve s systems of coupled 1D problems as in (4), with an $m_i N_h^i \times m_i N_h^i$ block matrix, whose factorization is stored once and for all at the first iteration and with $N_h^i < +\infty$ the dimension of the finite element space associated with $\Omega_{1D,i}$.

The choice of the modal multi-index $\mathbf{m}$ in (5) can be made a priori, as in [3], when we have some hints about u, or automatically, as in [4], if a suitable a posteriori modeling error estimator drives the selection of both Ω_i and $\mathbf{m}$.

2.3 Nodewise Hi-Mod Reduction

The piecewise Hi-Mod reduction represents a significative computational improvement with respect to the uniform Hi-Mod approach. Yet, it exhibits some limitations especially when dealing with extremely localized (almost pointwise) transverse dynamics or, on the contrary, with dynamics which involve the whole domain, even though with a different intensity (see Sect. 3 for an example). In the former case, a sufficiently large number of modes is assigned to a subdomain around the localized dynamic but, likely, the size of this domain will be excessively large compared with the entity of the dynamic; in the latter case, a piecewise Hi-Mod reduction may become uneffective so that the only feasible way is the uniform approach.

These considerations prompt us to set up a third Hi-Mod reduction procedure: the novelty is that now the modal functions are associated with the nodes of the finite element partition, in contrast to the piecewise approach where the modes are associated with subdomains of Ω. The association of the modes with the finite element nodes motivates the name chosen for this approach.

The trick which inspired us in setting up the nodewise approach consists of properly rewriting the modal expansion in the discrete space (3). By exploting the finite element basis $\{\vartheta_i\}$, we have indeed

$$v_m^h(x,y) = \sum_{k=1}^{m} \widetilde{v}_k^h(x)\,\varphi_k(\psi_x(y)) = \sum_{k=1}^{m} \Big[\sum_{i=1}^{N_h} \widetilde{v}_{k,i}^h\,\vartheta_i(x) \Big] \varphi_k(\psi_x(y)). \tag{7}$$

Notice that the leading role in such an expansion is taken by the summation on the modes. Simply by exchanging the two summations, we get

$$v_m^h(x,y) = \sum_{i=1}^{N_h} \Big[\sum_{k=1}^{m} \widetilde{v}_{k,i}^h\,\varphi_k(\psi_x(y)) \Big] \vartheta_i(x), \tag{8}$$

i.e., a representation for v_m^h, equivalent to (7), where the expansion runs over the finite element nodes. This leads us to define, in a straightforward way, a new discrete reduced space $V_{\mathbf{M}}^h$ where, ideally, the number of the modal basis functions may vary on each finite element node:

$$V_{\mathbf{M}}^h = \Big\{ v_{\mathbf{M}}^h(x,y) = \sum_{i=1}^{N_h} \Big[\sum_{k=1}^{m_i^N} \widetilde{v}_{k,i}^h\,\varphi_k(\psi_x(y)) \Big] \vartheta_i(x),\ \text{with } x \in \Omega_{1D},\ y \in \gamma_x \Big\}. \tag{9}$$

The global modal index m in (8) is here replaced by the nodewise modal index m_i^N, with $\mathbf{M} = \{m_i^N\}_{i=1}^{N_h}$ the vector of the modes for each finite element node.

The discrete nodewise Hi-Mod reduced formulation for (1) thus reads: for a certain modal multi-index $\mathbf{M} \in [\mathbb{N}^+]^{N_h}$,

$$\text{find } u_{\mathbf{M}}^h \in V_{\mathbf{M}}^h \quad : \quad a(u_{\mathbf{M}}^h, v_{\mathbf{M}}^h) = \mathscr{F}(v_{\mathbf{M}}^h) \quad \forall v_{\mathbf{M}}^h \in V_{\mathbf{M}}^h \tag{10}$$

where $a(\cdot,\cdot)$ and $\mathscr{F}(\cdot)$ coincides with the bilinear and linear forms in (1).

The algebraic counterpart of (10) is represented by a linear system whose matrix has a structure similar to that of the uniform case (with $m = \max_i m_i^N$), except that some rows and columns are deleted where $m_i^N < \max_i m_i^N$.

The change of perspective introduced by the nodewise Hi-Mod reduction relieves us from using a domain decomposition scheme in the presence of a different number of modal functions in Ω. This represents a significative improvement with respect to the piecewise Hi-Mod approach. No iterative procedure is now required to get the reduced solution; on the contrary, a domain decomposition scheme could now be employed to deal with more complex geometries (e.g., a bifurcation) not taken into account by the setting in Sect. 2.

The nodewise Hi-Mod reduction yields a reduced solution which is continuous, i.e., H^1-conformal, in Ω unlike the piecewise approach, where model discontinuities may occur. Moreover, the nodewise formulation makes sense, by definition, only after introducing the finite element basis. Spaces V_m^h and $V_{\mathbf{m}}^{b,h}$ have, on the contrary, a continuous counterpart obtained by replacing the modal coefficients in (3) and (6) with functions in V_{1D} and $H^1(\Omega_{1D}, \mathscr{T}_{\Omega_{1D}})$, respectively (see [1, 3] for the details).

Concerning the choice of the modal multi-index $\mathbf{M}$ in (10), we can ideally proceed via an a priori or an automatic selection, exactly as for the piecewise approach.

3 Numerical Assessment

We numerically validate the proposed Hi-Mod reduction procedures, to focus on the corresponding advantages and limits. In particular, we use affine finite elements to discretize the problem along Ω_{1D}, while employing sinusoidal functions to model the transverse dynamics. We evaluate the integrals of the sine functions via Gaussian quadrature formulas, based on, at least, four quadrature nodes per wavelenght.

First test case. This test case is meant to compare the three approaches onto the same full configuration. For this purpose, we consider a problem characterized by an analytical solution. We solve the Poisson problem $-\Delta u = f$ on $\Omega = (0,2) \times (0,\pi)$, completed with full homogeneous Dirichlet boundary conditions, so that $V \equiv H_0^1(\Omega)$, $V_{1D} \equiv H_0^1((0,2))$. The source term f is chosen such that the full solution is

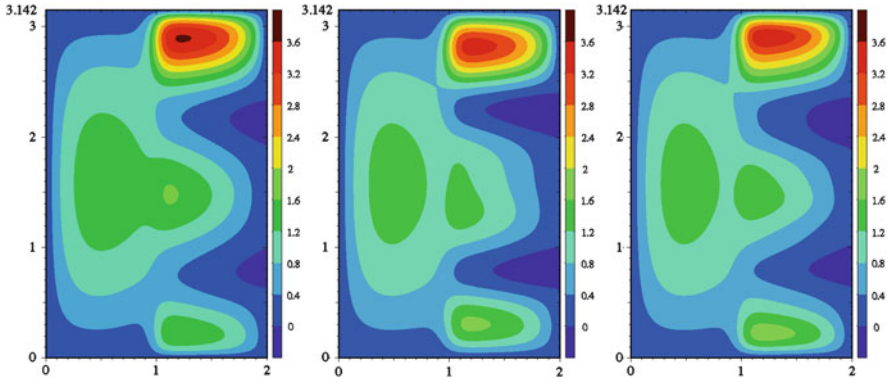


Fig. 1 First test case: full solution (*left*); uniformly reduced solutions, u_7^h (*center*), u_{16}^h (*right*)

$$u(x,y) = \frac{(256-x^8)(256-(2-x)^8)}{64800}\left\{\frac{100}{247}y(\pi-y)(2-x)\right.$$
$$\left. + y\left(\frac{\pi}{5}-y\right)\left(\frac{\pi}{3}-y\right)\left(\frac{3}{5}\pi-y\right)\left(\frac{3}{4}\pi-y\right)(\pi-y)\big(1+\tanh(10x-10)\big)\right\}.$$

In Fig. 1 (left) we show the contour plot of u approximated via a finite element scheme on a uniform unstructured grid of about 25,300 elements. Solution u clearly exhibits a smooth behaviour on the left part of Ω and a more irregular trend on the right.

We first apply the uniform Hi-Mod approach, by selecting $m = 7$ and $m = 16$ modes and choosing a uniform partition $\mathcal{T}_h$ of Ω_{1D} into 20 subintervals. Figure 1 (center-right) gathers the contour plots of the corresponding reduced solution: as expected, 16 modes provide us with a more close approximation, even though the difference between u_7^h and u_{16}^h is not so striking.

We successively assess the piecewise approach, inspired by the intrinsic heterogeneity of u. We split Ω into the subdomains $\Omega_1 = (0, 0.9) \times (0, \pi)$ and $\Omega_2 = (0.9, 2) \times (0, \pi)$; then we employ $m_1 = 1$ and $m_2 = 7$ modes, respectively and the same partition $\mathcal{T}_h$ as above. The domain decomposition algorithm (with relaxation equal to 0.5) converges after three iterations to the reduced solution $u_{1,7}^{b,h}$ in Fig. 2 (left). The model discontinuity is evident: we are in the presence of a nonconformal reduced solution. Formulation (6) guarantees indeed the continuity on Ω of both the trace and the flux only of the first $m_\perp^G = \min_{j=1}^s m_j$ modal components of $u_{\mathbf{m}}^b$ (i.e., only of the first one in such a case). More in general, as proved in [3], for a partition $\mathcal{T}_\Omega = \{\Omega_i\}_{i=1}^s$ of Ω, an H^1-conforming approximation is yielded only if $m_i > m_{i+1}$, for any $i = 1, \ldots, s-1$. By comparing Fig. 2 (left), e.g., with Fig. 1 (center), we recognize that a single mode is enough to describe u on Ω_1 with sufficient accuracy.

Finally, we resort to the node-wise Hi-Mod approach. The adopted modal distribution is shown in Fig. 2 (right): it is based on a uniform partition $\mathcal{T}_h$ with 50 subintervals. The corresponding reduced solution (see Fig. 2 (center)) is fully

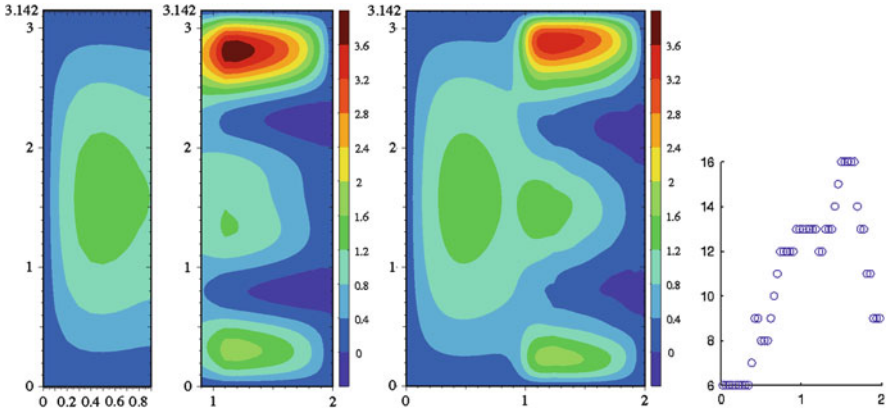


Fig. 2 First test case: piecewise reduced solution $u^{b,h}_{1,7}$ (*left*); nodewise reduced solution $u^h_{\mathbf{M}}$ (*center*) and corresponding modal ditribution (*right*)

comparable with the uniform one, u^h_{16}, in Fig. 1: nevertheless, 16 modes are now employed only on few nodes with a reduction of the size of the corresponding linear system, i.e., of the whole computational cost. As expected, the current reduced solution is continuous.

To make the comparison among the three approaches more quantitative, we consider the L^2-norm of the corresponding errors: $\|u - u^h_7\| = 0.2028$; $\|u - u^h_{16}\| = 0.0388$; $\|u - u^{b,h}_{1,7}\| = 0.2506$; $\|u - u^h_{\mathbf{M}}\| = 0.0566$. As expected, $u^{b,h}_{1,7}$ is thoroughly comparable with u^h_7, while the nodewise reduced solution is not so far from u^h_{16}.

Second test case. This test case provides an example of nodewise Hi-Mod reduction applied to a strong dynamic involving the whole Ω. We solve on $\Omega = (0, 4) \times (0, 1)$ the advection-diffusion problem $-\nabla \cdot \big(a(\mathbf{z})\nabla u\big) + \mathbf{b} \cdot \nabla u = 1$, with $a(\mathbf{z}) = 5 + 4.8 \sin(\pi x) \cos(\pi y)^{1/5}$ the diffusive coefficient, $\mathbf{b} = (100, 0)^T$ the advective field. We assign homogeneous Dirichlet boundary conditions along the horizontal sides, a nonhomogeneous Dirichlet datum, $u = 4 \sin(\pi x)$, at the inflow, homogeneous Neumann conditions at the outflow. This problem may model the density u of a fluid flowing horizontally (from left to right) in a media with a nonhomogenous permeability a. A distributed source, $f = 1$, is also present.

Due to the complex dynamics involved, it turns out to be a hard task to identify, a priori, suitable subdomains with a view to a piecewise approach. We consequently resort to both a uniform and a nodewise Hi-Mod reduction, by comparing the corresponding performances. Figure 3 (top) shows the uniform solution obtained by employing ten modes on the whole Ω. In Fig. 3 (bottom-left) we show the nodewise solution based on the modal distribution on the right. The two reduced solutions are really similar, but in the latter case at most eight modes are associated with a node. The order of the system reduces from 501 to 251.

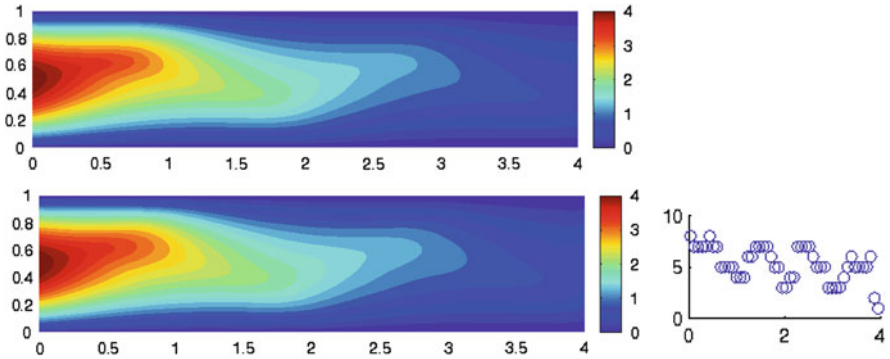


Fig. 3 Second test case: uniformly reduced solution u_{10}^h (*top*); nodewise reduced solution and associated modal distribution (*bottom*)

To summarize, the numerical assessment suggests that the nodewise Hi-Mod reduction is effective to deal with both localized and spread transverse dynamics.

References

1. Ern, A., Perotto, S., Veneziani, A.: Hierarchical model reduction for advection-diffusion-reaction problems. In: Kunisch, K., Of, G., Steinbach, O. (eds.) Numerical Mathematics and Advanced Applications, 703–710. Springer-Verlag (2008)
2. Formaggia, L., Nobile, F., Quarteroni, A., Veneziani, A.: Multiscale modelling of the circulatory system: a preliminary analysis. Comput. Visual. Sci. **2**, 75–83 (1999)
3. Perotto, S., Ern, A., Veneziani, A.: Hierarchical local model reduction for elliptic problems: a domain decomposition approach. Multiscale Model. Simul. **8**(4), 1102–1127 (2010)
4. Perotto, S., Ern, A., Veneziani, A.: Coupled adaptive model reduction and discretization for elliptic problems: a hierarchical approach. To be submitted (2012)